SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 1–20, January 1997

THE RESTARTED ARNOLDI METHOD APPLIED TO ITERATIVE LINEAR SYSTEM SOLVERS FOR THE COMPUTATION OF RIGHTMOST EIGENVALUES*

KARL MEERBERGEN† AND DIRK ROOSE†

Abstract. For the computation of a few eigenvalues of $Ax = \mu Bx$, the restarted Arnoldi method is often applied to transformations, e.g., the shift-invert transformation. Such transformations typically require the solution of linear systems. This paper presents an analysis of the application of the transformation $(M_A - \alpha M_B)^{-1}(A - \lambda B)$ to Arnoldi's method where α and λ are parameters and $M_A - \alpha M_B$ is some approximation to $A - \alpha B$. In fact, $(M_A - \alpha M_B)^{-1}$ corresponds to an iterative linear system solver for the system $(A - \alpha B)x = b$. The transformation is an alternative to the shift-invert transformation $(A-\alpha B)^{-1}B$ when direct system solvers are not available or not feasible. The restarted Arnoldi method is analyzed in the case of detection of the rightmost eigenvalues of real nonsymmetric matrices. The method is compared to Davidson's method by use of numerical examples.

Key words. Arnoldi's method, matrix transformations for eigenvalue problems, iterative linear system solvers and preconditioners

AMS subject classifications. 65F15, 65F50

PII. S0895479894274255

1. Introduction.

Consider the eigenvalue problem

$$Ax = \mu Bx, \tag{1}$$

where A and B are large sparse nonsymmetric real $N \times N$ matrices with eigenvalues $\mu_1, \mu_2, \ldots, \mu_N$ ordered by decreasing real part; i.e., $i > j \Rightarrow \mathrm{Re}(\mu_i) \leq \mathrm{Re}(\mu_j)$. The problem discussed in this paper is that of finding the rightmost eigenvalue(s) of (1) when N is large. The motivation of this work lies in the determination of the stability of the linearized system of the form

$$B\dot{x} = A(x), \quad x \in \mathbb{R}^N, \tag{2}$$

where x represents a state variable. A steady state solution x^* of such a nonlinear system is stable if the eigenvalues of (1) have negative real parts. We assume that N is large and A and B are sparse. The analysis in this paper applies to problems with any A and B, but we suppose that there is an α such that $A - \alpha B$ is nonsingular.

Popular methods for solving (1) are Krylov methods such as Arnoldi's method [20] or Lanczos's method [9] applied to the shift-invert transformation

$$T_{\mathrm{SI}} = (A - \alpha B)^{-1}B, \tag{3}$$

where $\alpha \in \mathbb{R}$ is called the shift. Since the eigenvalues of T_{SI} are $\theta_i = (\mu_i - \alpha)^{-1}$, $i = 1, \ldots, N$, the eigenvalues of (1) far from α correspond to eigenvalues of T_{SI} close to zero. The eigenvalues μ_i close to α are mapped to the well-separated extreme eigenvalues of T_{SI}. Typically, an eigenvalue solver applied to T_{SI} converges quickly

* Received by the editors September 14, 1994; accepted for publication (in revised form) by B. Kågström November 27, 1995.
http://www.siam.org/journals/simax/18-1/27425.html

† Department of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Heverlee-Leuven, Belgium (dirk.roose@cs.kuleuven.ac.be). Current address of the first author: Numerical Integration Technologies, Interleuvenlaan 70, B-3001 Heverlee-Leuven, Belgium (km@lmsnit.be).

to these well-separated extreme eigenvalues. Because (1) and T_{SI} have the same eigenvectors, it is feasible to recover eigenvalues and eigenvectors of (1) from the eigenvectors of T_{SI} (e.g., by use of the Rayleigh quotient). A Krylov method applied to T_{SI} requires the computation of several matrix–vector products $w = (A-\alpha B)^{-1}B\cdot v$, which involves solving the linear system

$$(A - \alpha B)w = Bv. \tag{4}$$

This is typically done by factorizing $A - \alpha B = LU$ and by using several back-substitutions of the form $w = U^{-1}(L^{-1}\, Bv)$ [6, 5, 18, 1, 10].

However, for large N, it may be advantageous to use iterative linear system solvers for computing the shift-invert transformation. This is discussed in section 2. We introduce the Cayley transform M_{C}, which allows the computation of eigenvalues of (1) close to a given λ using iterative linear system solvers. In section 3 we derive some spectral properties of M_{C}. The theory is valid for the generalized problem, but a few properties are restricted to the standard case $Ax = \mu x$. Section 4 analyzes the application of Arnoldi's method to M_{C}, and we present an asymptotic convergence result for the restarted Arnoldi method. We also briefly discuss Davidson's method. In section 5, numerical examples of the standard eigenvalue problem illustrate the theory. Finally, we conclude in section 6 with some general comments.

2. Shift-invert transformation with iterative linear system solvers. Arnoldi's method (see Algorithm 1 in section 4.2) applied to T_{SI} computes the Krylov space

$$\mathcal{K}_m(T_{\mathrm{SI}}, v_1) = \mathrm{span}\{v_1, T_{\mathrm{SI}}v_1, T_{\mathrm{SI}}^2 v_1, \ldots, T_{\mathrm{SI}}^{m-1}v_1\}.$$

The computation of $\mathcal{K}_m$ requires several matrix–vector products $w = (A-\alpha B)^{-1}B\cdot v$. In this section we discuss whether iterative linear system solvers are suitable for building $\mathcal{K}_m(T_{\mathrm{SI}}, v_1)$.

A straightforward approach to computing w is to apply a Krylov linear system solver like GMRES [23], BICGSTAB(ℓ) [26], or QMR [7] to (4). This technique was used by Mittelmann et al. [12, 11] for computing eigenvalues by inverse iteration using LSQR, SYMMLQ, and GMRES. In general, it is better to solve iteratively the preconditioned system

$$\hat{P}^{-1}(A - \alpha B)w = \hat{P}^{-1}(Bv)$$

by a Krylov method where the preconditioner $\hat{P}^{-1}$ is a good approximation to $(A - \alpha B)^{-1}$.

In this paper, however, the linear system solver is supposed to be stationary (see also [2]); i.e., the approximate solution of (4) can be written as

$$\hat{w} = Gw_0 + (M_A - \alpha M_B)^{-1}Bv, \tag{5}$$

where w_0 is the initial solution, $M_A - \alpha M_B$ is an approximation to $A - \alpha B$, and

$$G = I - (M_A - \alpha M_B)^{-1}(A - \alpha B) \tag{6}$$

is called the iteration matrix. Typical stationary solvers are Jacobi and Gauss–Seidel-type relaxation methods, multigrid solvers, and incomplete LU factorizations like ILU

or ILUT [22] (ILU with threshold) or ILUTP [22] (ILUT with pivoting). For all these solvers, the matrix

$$M_{\mathrm{SI}} = (M_A - \alpha M_B)^{-1} B \tag{7}$$

provides an approximation to T_{SI}. Note that $M_A - \alpha M_B$ is just a notation for an approximation to $A - \alpha B$. The matrices M_A and M_B cannot be viewed as approximations to A and B, respectively. There is a problem when (5) is used for computing $\mathcal{K}_m$. Suppose that $w_0 = 0$; then (5) becomes $\hat{w} = M_{\mathrm{SI}} v$ for each $v \in \mathbb{C}^N$. So, in Arnoldi's method, the Krylov space $\mathcal{K}_m(M_{\mathrm{SI}}, v_1)$ rather than $\mathcal{K}_m(T_{\mathrm{SI}}, v_1)$ is computed. In fact, the eigenvectors and eigenvalues of M_{SI} are computed. Since the rightmost eigenvalues (and associated eigenvectors) of (1) are recovered from the eigenvectors computed by Arnoldi's method, the eigenvectors of M_{SI} and of $Ax = \mu Bx$ should lie close to each other. To achieve this, the linear system (4) must be solved accurately, which can be very expensive.

However, we can take advantage of the fact that we are interested only in the rightmost eigenvalues of (1); thus only a few eigenvectors of M_{SI} must be very good approximate eigenvectors of (1). We suggest the use of

$$M_{\mathrm{C}} = (M_A - \alpha M_B)^{-1} (A - \lambda B) \tag{8}$$

in Arnoldi's method and Lanczos's method instead of M_{SI}. Here λ is an approximation to the rightmost eigenvalue μ_1 of $Ax = \mu Bx$. In section 3, we show the relation with the Cayley transform $T_{\mathrm{C}} = (A - \alpha B)^{-1}(A - \lambda B)$. The link with shift-invert Arnoldi is then clear, since Arnoldi's method applied to T_{C} produces in exact arithmetic the same results as T_{SI} [10]. To understand the transformation M_{C}, we consider the following example.

Example 1. Consider the standard eigenvalue problem with

$$A = \begin{bmatrix} -2 & 1 & 0 & 0 \\ 1 & -3 & 1 & 0 \\ 0 & 1 & -4 & 1 \\ 0 & 0 & 1 & -5 \end{bmatrix}.$$

The eigenvalues μ_i and eigenvectors x_i, $i = 1, \ldots, 4$ are

$$[\mu_i]_{i=1}^4 = \begin{bmatrix} -1.2547 \\ -2.8227 \\ -4.1773 \\ -5.745 \end{bmatrix}$$

and

$$[x_1, \ldots, x_4] = \begin{bmatrix} 0.7780 & 0.5533 & 0.2912 & 0.0625 \\ 0.5798 & -0.4552 & -0.6340 & -0.2339 \\ 0.2339 & -0.6340 & 0.4552 & 0.5798 \\ 0.0625 & -0.2912 & 0.5533 & -0.7780 \end{bmatrix}.$$

Let $\alpha = 0$. We use one iteration of the Jacobi method with $w_0 = 0$ for solving (4), so $M_{\mathrm{SI}} = M_A^{-1} = [\mathrm{diag}(A)]^{-1} = \mathrm{diag}(-0.5, -0.333, -0.25, -0.2)$. The eigenvectors of M_A^{-1} are the columns of the unit matrix and differ greatly from the eigenvectors of

A. Hence, Arnoldi's method applied to M_{SI} cannot be used to recover eigenvalues of A. For $\lambda = \mu_1$, the matrix $M_{\mathrm{C}} = [\mathrm{diag}(A)]^{-1}(A - \lambda I)$ is

$$M_{\mathrm{C}} = \begin{bmatrix} 0.3726 & -0.5000 & 0 & 0 \\ -0.3333 & 0.5818 & -0.3333 & 0 \\ 0 & -0.2500 & 0.6863 & -0.2500 \\ 0 & 0 & -0.2000 & 0.7491 \end{bmatrix}.$$

$(\lambda = \mu_1, x_1)$ is an eigenpair of A, so $(A - \lambda I)x_1 = 0$ and hence $(0, x_1)$ is an eigenpair of M_{C}. The eigenvalues η_i and the (normalized) eigenvectors $u_1, \ldots, u_4$ of M_{C} are

$$[\eta_i]_{i=1}^4 = \begin{bmatrix} 0 \\ 1.1064 \\ 0.8175 \\ 0.4659 \end{bmatrix}, \qquad [u_1, \ldots, u_4] = \begin{bmatrix} 0.7780 & 0.4301 & -0.5675 & -0.6047 \\ 0.5798 & -0.6312 & 0.5049 & 0.1128 \\ 0.2339 & 0.5633 & 0.2105 & 0.6439 \\ 0.0625 & -0.3153 & -0.6153 & 0.4549 \end{bmatrix}.$$

It can be seen that $x_1 = u_1$, so μ_1 can be computed as $\mu_1 = u_1^H A u_1$. The other eigenvectors of M_{C} differ a lot from those of A, which makes recovering μ_2, μ_3, or μ_4 from u_2, u_3, or u_4 impossible.

In this example, $\lambda = \mu_1$ is used, but in practice μ_1 is unknown, so a practical algorithm will vary the parameter λ until $\lambda = \mu_1$; see sections 4 and 5. This transformation was presented first for use in Davidson's method by Morgan and Scott [14], Morgan [13], Sadkane [24], and Crouzeix, Philippe, and Sadkane [4], often with $\alpha = \lambda$ (see also Remark 3.2 at the end of section 3), and in the Jacobi–Davidson method by Sleijpen and Van der Vorst [27]. A similar transformation was suggested for Lanczos's method by Morgan and Scott [15].

3. Spectral properties of M_{C}. In this section, we derive relations between the eigenpairs of M_{C} and T_{C}. Note that the analysis in this section does not relate to the convergence of Arnoldi's method applied to M_{C}. We denote the eigenpairs of M_{C} and T_{C} by (η_k, u_k) and (θ_i, x_i), respectively. The following lemma forms the basis for the analysis of this section.

LEMMA 3.1. *The transformation M_{C} consists of two parts:*

$$M_{\mathrm{C}} = T_{\mathrm{C}} - G T_{\mathrm{C}} \tag{9}$$

with $T_{\mathrm{C}} = (A - \alpha B)^{-1}(A - \lambda B)$ and with G given by (6).

Proof.

$$\begin{aligned} M_{\mathrm{C}} &= (M_A - \alpha M_B)^{-1}(A - \lambda B) \\ &= [(M_A - \alpha M_B)^{-1}(A - \alpha B)] \cdot [(A - \alpha B)^{-1}(A - \lambda B)] \\ &= [I - G] \cdot [T_{\mathrm{C}}] \\ &= T_{\mathrm{C}} - G T_{\mathrm{C}}. \end{aligned}$$

∎

The matrix G provides a measure for the deviation of M_{SI} and T_{SI}. When $\|G\|$ is small, $(M_A - \alpha M_B)^{-1}$ represents a good linear system solver. The matrix M_{C} can be viewed as a perturbation of the matrix T_{C}. Perturbation analysis and a posteriori error analysis [28, 3, 21] can help find bounds on the eigenvalues and eigenvectors of M_{C}. However, the classical bounds for $T = M + tE$ are derived in function of a norm of E and t. They are first-order approximations in t, assuming t is small. In this application, the perturbation tE is not always small and has a special form. Therefore, we derive bounds for this specific problem. In the following analysis, the

iterative linear system solver is supposed to be stationary; see section 2. Theorem 3.3 gives a bound on the eigenvalues of M_{C}. Theorem 3.4 provides a bound on the eigenvectors of M_{C}.

First, we define the (spectral) projector $\mathcal{P}_i$ such that

$$\begin{aligned} \mathcal{P}_i x_i &= x_i, \\ \mathcal{P}_i x_j &= 0, \ i \neq j \end{aligned} \tag{10}$$

and the complementary projector (or complement) $\mathcal{Q}_i = I - \mathcal{P}_i$ such that

$$\begin{aligned} \mathcal{Q}_i x_i &= 0, \\ \mathcal{Q}_i x_j &= x_j, \ i \neq j. \end{aligned} \tag{11}$$

If x_i and y_i are the right and left eigenvectors of T_{C} corresponding to θ_i, then $y_i^H x_j = 0$ when $j \neq i$ and y_i can be scaled such that $y_i^H x_i = 1$. Hence $\mathcal{P}_i = x_i y_i^H$ and $\mathcal{Q}_i = I - x_i y_i^H$. We also recall the following lemma.

LEMMA 3.2 (Bauer–Fike [21, 3]). *Consider a matrix $T \in \mathbb{C}^{N\times N}$ with (simple) eigenvalues $\theta_1, \ldots, \theta_N$ and eigenvectors denoted by the matrix $X = [x_1, \ldots, x_N]$. Given an approximate eigenpair (σ, s) with $\|s\|_2 = 1$ for which the residual*

$$Ts - \sigma s = e, \tag{12}$$

there is an eigenvalue θ_k of T such that

$$|\theta_k - \sigma| \leq \mathrm{cond}_2(X)\, \|e\|_2$$

with $\mathrm{cond}_2(X) = \|X\|_2 \|X^{-1}\|_2$.

There also exist similar formulae for multiple eigenvalues. In our analysis, we suppose M_{C} to have N simple eigenvalues.

THEOREM 3.3. *Suppose that M_{C} has N simple eigenvalues $\eta_1, \ldots, \eta_N$ and that the eigenvectors are denoted by the matrix $U = [u_1, \ldots, u_N]$. For each eigenvalue θ_i of T_{C}, $i = 1, \ldots, N$, there is an eigenvalue η_k of M_{C}, such that*

$$|\theta_i - \eta_k| \leq |\theta_i|\, \mathrm{cond}_2(U)\, \|G\mathcal{P}_i\|_2.$$

Proof. From (9), it follows that

$$\begin{aligned} (M_{\mathrm{C}} - T_{\mathrm{C}})x_i &= -GT_{\mathrm{C}}x_i, \\ M_{\mathrm{C}}x_i - \theta_i x_i &= -\theta_i G x_i \\ &= -\theta_i G\mathcal{P}_i x_i. \end{aligned}$$

By applying Lemma 3.2 with $\|x_i\|_2 = 1$, we see that there is an η_k, $1 \leq k \leq N$ such that

$$|\eta_k - \theta_i| \leq |\theta_i|\, \mathrm{cond}_2(U)\, \|G\mathcal{P}_i\|_2. \qquad \square$$

From the definition of M_{C}, it can be seen that the eigenvectors x_i of T_{C} corresponding to θ_i close to zero are nearly eigenvectors of M_{C}. However, eigenvectors of M_{C} are not necessarily approximate eigenvectors of T_{C}. For the link between x_i and u_k, we prove Theorem 3.4.

THEOREM 3.4. *Suppose that M_{C} has distinct eigenvalues. Consider an eigenvector x_i of T_{C} with associate eigenvalue θ_i and let $\|x_i\|_1 = 1$. Then for each $i = 1, \ldots, N$ there is an eigenpair (η_k, u_k) of M_{C} with $x_i = \omega_k u_k - e$, $\|u_k\|_1 = 1$, and*

$$\|e\|_1 \leq 2N\mathrm{cond}_1(U)\frac{|\theta_i|\,\|\mathcal{Q}_i G\mathcal{P}_i\|_1}{\min_{j\neq k}|\eta_j - \eta_k|}, \tag{13}$$

$$|\omega_k| \geq 1 - \|e\|_1. \tag{14}$$

A consequence is that when $\|e\|_1$ is small, $|\omega_k| \simeq 1$.

Proof. Using the spectral projectors $\mathcal{P}_i$ and $\mathcal{Q}_i$ defined by (10) and (11), M_{C} can be written as

$$\begin{aligned} M_{\mathrm{C}} &= T_{\mathrm{C}} - \mathcal{P}_i G T_{\mathrm{C}} - \mathcal{Q}_i G T_{\mathrm{C}} \\ &= (I - \mathcal{P}_i G) T_{\mathrm{C}} - \mathcal{Q}_i G T_{\mathrm{C}}. \end{aligned}$$

From the definition of $\mathcal{P}_i$, it turns out that there is a $\gamma_i \in \mathbb{C}$ such that $(I - \mathcal{P}_i G)x_i = \gamma_i x_i$. As a result,

$$(I - \mathcal{P}_i G)T_{\mathrm{C}} x_i = \gamma_i \theta_i x_i. \tag{15}$$

Now define ϵ such that $\eta_k = \gamma_i\theta_i + \epsilon$. Let $x_i = \sum_{j=1}^N \omega_j u_j$, $\|u_j\|_1 = 1$, and so, $e = -\sum_{j=1, j\neq k}^N \omega_j u_j$. Since $M_{\mathrm{C}}(x_i + e) = \eta_k(x_i + e)$, it follows that

$$(M_{\mathrm{C}} - \eta_k I)e = \eta_k x_i - M_{\mathrm{C}} x_i.$$

From $\eta_k x_i = \gamma_i\theta_i x_i + \epsilon x_i$ and (15), it follows that $\eta_k x_i = (I - \mathcal{P}_i G)T_{\mathrm{C}} x_i + \epsilon x_i$ and

$$\begin{aligned} (M_{\mathrm{C}} - \eta_k I)e &= \epsilon x_i + ((I - \mathcal{P}_i G)T_{\mathrm{C}} - M_{\mathrm{C}})x_i \\ &= \epsilon x_i + \theta_i \mathcal{Q}_i G x_i. \end{aligned} \tag{16}$$

Denote the left eigenvector of M_{C} associated with η_j by s_j scaled such that $s_j^H u_j = 1$. Multiply (16) by s_j^H; then we have

$$s_j^H (M_{\mathrm{C}} - \eta_k I)e = \epsilon s_j^H x_i + \theta_i s_j^H \mathcal{Q}_i G x_i. \tag{17}$$

Since $s_j^H M_{\mathrm{C}} = \eta_j s_j^H$ and, for $j \neq k$, $s_j^H e = -\omega_j$, we have

$$-(\eta_j - \eta_k)\omega_j = \epsilon s_j^H x_i + \theta_i s_j^H \mathcal{Q}_i G x_i. \tag{18}$$

For $j = k$, (17) becomes

$$0 = \epsilon s_k^H x_i + \theta_i s_k^H \mathcal{Q}_i G x_i.$$

We choose k such that $|s_j^H x_i|$ is maximal for $j = k$, so $s_k^H x_i \neq 0$, and hence

$$\epsilon = -\theta_i \frac{s_k^H \mathcal{Q}_i G x_i}{s_k^H x_i} \tag{19}$$

exists. Since $\eta_j \neq \eta_k$ for $k \neq j$, and by combining (18) and (19), we find that

$$\omega_j = \frac{\theta_i}{\eta_k - \eta_j}\left(s_j^H - \frac{s_j^H x_i}{s_k^H x_i} s_k^H\right)\mathcal{Q}_i G s_i.$$

Note that k is chosen such that $|s_j^H x_i|/|s_k^H x_i| \leq 1$ and note also that $Gx_i = G\mathcal{P}_i x_i$. Hence,

$$|\omega_j| \leq \frac{|\theta_i|}{|\eta_j - \eta_k|} 2 \max_j(\|s_j\|) \|\mathcal{Q}_i G \mathcal{P}_i\| \, \|x_i\|.$$

Recall that $\|U\|_1 = \max_j(\|u_j\|_1) = 1$ and $\|U^{-1}\|_1 = \max_j(\|s_j\|_1)$. Thus,

$$|\omega_j| \leq 2 \frac{|\theta_i|}{|\eta_j - \eta_k|} \|U\|_1 \|U^{-1}\|_1 \|\mathcal{Q}_i G \mathcal{P}_i\|_1 \|x_i\|_1$$

for $j = 1, \ldots, k-1, k+1, \ldots, N$. From $\|e\|_1 \leq \sum_{j \neq k} |\omega_j|$ (13) follows. Equation (14) follows from $\|x_i\|_1 \leq |\omega_k| \|u_k\|_1 + \|e\|_1$. ∎

From (13) it follows that there is an eigenvector u_k of M_C that approximates the eigenvector x_i well if (a) $\|\mathcal{Q}_i G \mathcal{P}_i\|_1$ is small, (b) $|\theta_i|$ is small, and (c) η_k is a well-separated eigenvalue of M_C. These three conditions play an important role in the mapping of an eigenvector u_k of M_C to an eigenvector x_i of T_C. In the following paragraphs, each of these conditions is discussed.

(a) Small $\|\mathcal{Q}_i G \mathcal{P}_i\|_1$. The norm $\|\mathcal{Q}_i G \mathcal{P}_i\|_1$ is a measure of the portion of x_j in Gx_i, $j \neq i$. It depends on the linear system solver used, the shift α, and the eigenstructure of (1). A general analysis of all these parameters is hard to do, but a few general properties can be given. We restrict the analysis in this paragraph to the standard case $B = I$. (The extension to the generalized case is not obvious, but some properties are easily transferred.) In the standard case, G may be written as

$$G = I - (M_A - \alpha I)^{-1}(A - \alpha I).$$

We demonstrate the influence of α on $\|G\|$ and $\|\mathcal{Q}_i G \mathcal{P}_i\|$ by Lemma 3.5 and Theorem 3.6.

LEMMA 3.5. *Define L by $A - \alpha I = M_A - \alpha I + L$. Let $(A - \alpha I)^{-1}$ be diagonalizable and XZX^{-1} be its Jordan canonical form. If $\|(A - \alpha I)^{-1} L\|_2 = \epsilon < 1$, then*

$$\|G\|_2 \leq \frac{\epsilon}{1 - \epsilon}.$$

Moreover,

$$\epsilon \leq \frac{\|L\|_2}{\min_j |\mu_j - \alpha|} \mathrm{cond}_2(X).$$

Proof. From (6) and $A - \alpha I = M_A - \alpha I + L$,

$$\begin{aligned} G &= I - (M_A - \alpha I)^{-1}(A - \alpha I), \\ (M_A - \alpha I)G &= (M_A - \alpha I) - (A - \alpha I), \\ (A - \alpha I - L)G &= -L, \\ (I - (A - \alpha I)^{-1} L)G &= (A - \alpha I)^{-1} L, \\ G &= (I - (A - \alpha I)^{-1} L)^{-1} (A - \alpha I)^{-1} L. \end{aligned} \tag{20}$$

Hence

$$\|G\|_2 \leq \|(I - (A - \alpha I)^{-1} L)^{-1}\|_2 \|(A - \alpha I)^{-1} L\|_2.$$

From Lemma 2.3.3 in Golub and Van Loan [8, p. 59] it follows that

$$\|(I-(A-\alpha I)^{-1}L)^{-1}\|_2 \leq \frac{1}{1-\|(A-\alpha I)^{-1}L\|_2}$$

if $\|(A-\alpha I)^{-1}L\|_2 < 1$. This shows the first part of the lemma.

Let us now prove the second part. By using the Jordan canonical form of $(A-\alpha I)^{-1}$, we find that

$$(A-\alpha I)^{-1}L = XZX^{-1}L.$$

Hence

$$\|(A-\alpha I)^{-1}L\|_2 \leq \|X\|_2\|Z\|_2\|X^{-1}\|_2\|L\|_2,$$

from which the second part of the lemma follows. ∎

A similar property is shown for $\|\mathcal{Q}_i G\mathcal{P}_i\|$. (An analogous result was proven by Morgan [13, Theorem 2].)

THEOREM 3.6. *Under the conditions and definitions in Lemma* 3.5*,*

$$\|\mathcal{Q}_i G\mathcal{P}_i\|_2 \leq \frac{\|L\|_2}{\min_{j\neq i}|\mu_j-\alpha|}\mathrm{cond}_2(X)\frac{1}{1-\epsilon}\|\mathcal{P}_i\|_2.$$

Proof. From (20),

$$\begin{aligned}(A-\alpha I)G &= LG-L,\\ \mathcal{Q}_i G\mathcal{P}_i &= \mathcal{Q}_i(A-\alpha I)^{-1}L(G-I)\mathcal{P}_i.\end{aligned}$$

Hence

$$\begin{aligned}\|\mathcal{Q}_i G\mathcal{P}_i\|_2 &\leq \|\mathcal{Q}_i(A-\alpha I)^{-1}\|_2\|L\|_2\|\mathcal{P}_i\|_2(1+\|G\|_2)\\ &\leq \frac{\mathrm{cond}_2(X)}{\min_{j\neq i}|\mu_j-\alpha|}\|L\|_2\|\mathcal{P}_i\|_2\left(1+\frac{\epsilon}{1-\epsilon}\right).\end{aligned}$$ ∎

Suppose that L is independent of α, which is the case in iterative linear system solvers such as Jacobi and Gauss–Seidel. (Note that this is not the case in general, e.g., for incomplete factorizations.) Then, following Theorem 3.6, $\|\mathcal{Q}_i G\mathcal{P}_i\|$ can be very small when the eigenvalues μ_j, $j \neq i$ lie far from α, even when μ_i lies close to α. Unfortunately, this implies that μ_i should be a well-separated eigenvalue of $Ax = \mu Bx$ and this is, in general, not the case. Therefore, the only practical way to reduce $\|\mathcal{Q}_i G\mathcal{P}_i\|$ is to move α away from the spectrum. We illustrate this in Example 2. Since, in general, L depends on α, this can only be viewed as a qualitative statement.

Example 2. The Olmstead model [16] represents the flow of a layer of viscoelastic fluid heated from below. The equations are

$$\begin{cases}\dfrac{\partial u}{\partial t} = (1-C)\dfrac{\partial^2 v}{\partial X^2} + C\dfrac{\partial^2 u}{\partial X^2} + Ru - u^3,\\ B\dfrac{\partial v}{\partial t} = u - v\end{cases}$$

with boundary conditions $u(0) = u(1) = 0$ and $v(0) = v(1) = 0$. u represents the speed of the fluid and v is related to viscoelastic forces. The equation was discretized with central differences with grid size $h = 1/(N/2)$. After discretization, the equation

may be written as $\dot{x} = f(x)$ with $x = [u_1, v_1, u_2, v_2, \ldots, u_{N/2}, v_{N/2}]^T$. For the parameter values $B = 2$, $C = 0.1$, and $R = 4.7$, the equation has the trivial steady state solution $[u, v] = 0$. The size of the Jacobian matrix $A = \partial f/\partial x$ is 1000, and the rightmost eigenvalue of $Ax = \lambda x$ is $\mu_1 = 4.510184$. Linear systems with $A - \alpha I$ were solved approximately by the Sparskit routine ILUT(lfil=2, tol=0.001) [22]. Table 1 shows $\|G\|$ and $\|\mathcal{Q}_1 G \mathcal{P}_1\|$ for three values of α.

TABLE 1

α	$\|G\|_1$	$\|\mathcal{Q}_1 G \mathcal{P}_1\|_1$
4.6	$2\,10^{6}$	$5\,10^{2}$
6	$3\,10^{-1}$	$3\,10^{-6}$
10	$1\,10^{-12}$	$2\,10^{-14}$

It is clear that shifting α to the right makes $\|G\|$ and $\|\mathcal{Q}_1 G \mathcal{P}_1\|$ smaller. It is also clear in this example that $\|\mathcal{Q}_1 G \mathcal{P}_1\| \ll \|G\|$ if α lies close to μ_1.

(b) Small $|\theta_i|$. Recall that $T_{\mathrm{C}} = (A - \alpha B)^{-1}(A - \lambda B)$; hence the transformation T_{C} has eigenvalues

$$\theta_i = \frac{\mu_i - \lambda}{\mu_i - \alpha} = 1 + (\alpha - \lambda)\frac{1}{\mu_i - \alpha}. \tag{21}$$

In fact, the θ_i are given by the Cayley transform or Möbius transform [25] of μ_i. $|\theta_i|$ is small when

$$\left|\frac{\mu_i - \lambda}{\mu_i - \alpha}\right| \leq \epsilon \tag{22}$$

for a given positive ϵ. Formula (22) holds for μ_i lying in the disk with center $(\epsilon^2\alpha - \lambda)/(\epsilon^2 - 1)$ and radius $|\alpha - \lambda|\epsilon/|1 - \epsilon^2|$ [10]. When ϵ is small, the center is approximately λ and the radius is $|\alpha - \lambda|\epsilon$. Also note that when α moves away from λ, the radius of the circle becomes larger. This means that the region of eigenvalues in which the associated eigenvectors correspond well to those of M_{C} becomes larger.

(c) Well-separated eigenvalues. Recall that μ_1 is the rightmost eigenvalue. Suppose that $\lambda = \mu_1$. Then T_{C} maps the rightmost eigenvalue μ_1 to $\theta_1 = 0$. If α lies to the right of λ, the other eigenvalues lie relatively far from α and are mapped relatively close to 1 (see the second expression in (21)). A typical situation is shown in Figure 1. The "gap" σ between θ_1 and the cluster of eigenvalues close to 1 becomes larger when α moves to λ. If $\|G\|$ is small, then the spectrum of M_{C} is similar in shape to the spectrum of T_{C}. Hence the eigenvalues of M_{C} are well separated when α lies close to λ (see Figure 1).

It appears that α plays a conflicting role concerning the mapping from $Ax = \mu Bx$ to $M_{\mathrm{C}}u = \eta u$: when α moves further to the right of μ_1, we see that

(i) $\|\mathcal{Q}_1 G \mathcal{P}_1\|$ is smaller, which is good,

(ii) the disk (22) becomes larger, which is also good,

(iii) the eigenvalues of T_{C} become less separated, which implies less separation of the eigenvalues of M_{C}, which is bad.

Thus the conditions are difficult to reconcile.

Remark 3.1. Note that in the analysis above the parameter λ may take arbitrary values. If λ is a simple eigenvalue of $Ax = \mu Bx$, say $\lambda = \mu_1$, the bounds in Theorems 3.3 and 3.4 can be made sharper. Since $\theta_1 = 0$ and $\theta_i \neq 0$ for $i > 1$, we have $T_{\mathrm{C}} = \mathcal{Q}_1 T_{\mathrm{C}}$. Hence

$$M_{\mathrm{C}} = T_{\mathrm{C}} - GT_{\mathrm{C}} = T_{\mathrm{C}} - G\mathcal{Q}_1 T_{\mathrm{C}}.$$

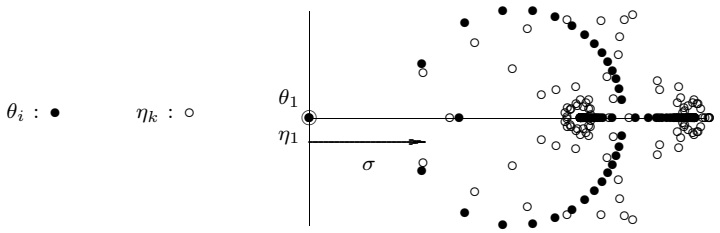


FIG. 1. *Spectrum of T_C and M_C for $\alpha = 5$ and $\lambda = \mu_1 \simeq 1.695$ for the Olmstead model for $R = 2$ and $N = 100$. M_C is formed by 10 Gauss–Seidel iterations.*

If the matrix T_C is normal, then $\mathcal{Q}_1$ is an orthogonal projector, so $\|\mathcal{Q}_1\|_2 = 1$. As a result $\|G\mathcal{Q}_1\|_2 \leq \|G\|_2$. If T_C is slightly nonnormal $\|G\mathcal{Q}_1\| < \|G\|$ is often valid, but if T_C is highly nonnormal, i.e., when left and right eigenvectors y_1 and x_1 are nearly orthogonal, then $\|G\mathcal{Q}_1\| > \|G\|$ is possible. By replacing G by $G\mathcal{Q}_1$ in Theorems 3.3 and 3.4, the bounds on $\|x_i - u_k\|$ and $|\theta_i - \eta_k|$ are sharper for normal and slightly nonnormal matrices when λ is an eigenvalue of $Ax = \mu Bx$.

Remark 3.2. In Davidson's method the choice of $\alpha = \lambda$ is very popular. If $\lambda = \alpha = \mu_1$, then

$$M_C = (M_A - \lambda M_B)^{-1}(A - \lambda B) = I - G.$$

Because $(0, x_1)$ is an eigenpair of M_C,

$$M_C = \mathcal{Q}_1 - G\mathcal{Q}_1.$$

The theorems above can still be used with $T_C = \mathcal{Q}_1$. For the details of this choice of α, see the work of Morgan [13] and Crouzeix, Philippe, and Sadkane [4].

4. Algorithms.

4.1. Implementation of M_C. A black box iterative linear system solver applied to the linear system $(A - \alpha B)z = b$ improves an initial guess z_0 by

$$z_1 = Gz_0 + (M_A - \alpha M_B)^{-1}b \tag{23}$$

with G given by (6). G can be viewed as the iteration matrix of a solver consisting of one iteration of (23). The input of the black box iterative solver consists of the matrix $(A - \alpha B)$, the initial solution z_0, and the right-hand side b. The multiplication $M_C \cdot v$ can be written as follows:

$$w = M_C v = (M_A - \alpha M_B)^{-1}(A - \lambda B)v. \tag{24}$$

Expression (24) is computed from (23) with $z_0 = 0$, $b = (A - \lambda B)v$, and $w = z_1$. The calculation of $b = (A - \lambda B)v$ requires two matrix–vector products. The multiplication by A can be avoided as follows. Rewrite (24) as

$$\begin{aligned} w = M_C v &= (M_A - \alpha M_B)^{-1}(A - \alpha B)v + (M_A - \alpha M_B)^{-1}(\alpha - \lambda)Bv \\ &= (I - G)v + (M_A - \alpha M_B)^{-1}(\alpha - \lambda)Bv \\ &= v - [Gv - (M_A - \alpha M_B)^{-1}(\alpha - \lambda)Bv]. \end{aligned} \tag{25}$$

Expression (25) is computed as $w = v - z_1$, where z_1 results from (23) with $b = -(\alpha - \lambda)Bv$ and $z_0 = v$. Note that in Arnoldi's process $w = z_1 = (I - M_C)v$ can be used rather than $w = v - z_1$ such that the first term in the right-hand side of (25) can be thrown away, since M_C and $I - M_C$ give the same Krylov basis.

4.2. The restarted Arnoldi method. Arnoldi's process (inner loop in Algorithm 1) computes an orthonormal basis $\{v_1, \ldots, v_m\}$ of the Krylov space $\mathcal{K}_m = \mathcal{K}_m(M_C, v_1)$. It is well known that $\mathcal{K}_m$ is rich in eigenvectors of M_C associated with the well-separated extreme eigenvalues of M_C [21]. If α is chosen as explained at the end of section 3, then the eigenvalues of M_C close to zero are well separated. Hence $\mathcal{K}_m$ is rich in the eigenvectors x_i associated with eigenvalues of $Ax = \mu Bx$ close to λ. An approximate eigenpair $(\hat{\mu}_i, \hat{x}_i)$ is recovered from the eigenpairs of $H_m z = \mu F_m z$ with $H_m = V_m^H A V_m$ and $F_m = V_m^H B V_m$ by the QZ method; see Algorithm 1. If $B = I$, then the eigenvalues are computed from $H_m z = \mu z$ by the QR method. In each outer iteration, i.e., after each Arnoldi process, the parameters λ and α are reset and the process is restarted. The approximate eigenvector $\hat{x}_1$ is used as initial vector v_1. After each outer iteration, the convergence is tested by use of the residual norm $\|A\hat{x}_1 - \hat{\mu}_1 B\hat{x}_1\|_2 / \|\hat{x}_1\|_2$.

ALGORITHM 1 (the restarted Arnoldi method applied to M_C).
Given α, λ, and v_1 with $\|v_1\|_2$=1.
repeat
  Set up the linear system solver for $(A - \alpha B)w = Bv$.
  for $i = 1$ **to** $m - 1$ **do**
    Form $w_i = M_C v_i$.
    Orthonormalize w_i against $v_1, \ldots, v_i$.
    Let $v_{i+1} = w_i$.
  end for
  Compute $H_m = V_m^H A V_m$ and $F_m = V_m^H B V_m \in \mathbb{C}^{m \times m}$ with $V_m = [v_1, \ldots, v_m]$.
  Compute the eigenpairs $(\hat{\mu}_j, \hat{z}_j)$ of $H_m z = \mu F_m z$ by the QZ method.
  Form $\hat{x}_1 = V_m \hat{z}_1$.
  Set $\lambda = \hat{\mu}_1$ and reset α (e.g., $\alpha = \mathrm{Re}(\lambda)$).
  Set $v_1 = \hat{x}_1 / \|\hat{x}_1\|_2$.
until $\|A\hat{x}_1 - \hat{\mu}_1 B\hat{x}_1\|_2 < \epsilon \|\hat{x}_1\|_2$

Before we analyze the convergence of Arnoldi's method, we consider the following example.

Example 3. Recall the Olmstead model from Example 2. Here we consider the matrix of size $N = 100$ for the parameter value $R = 2$. The rightmost eigenvalue of A is the real value $\mu_1 = 1.6905$. One iteration of Algorithm 1 was run with $m = 20$, $\lambda = \mu_1 = 1.6905$, and α, as shown in Table 2. Of course, since μ_1 is unknown, this can never be done in practice, but the results give us interesting information on the convergence of Algorithm 1. The linear systems were solved approximately by 10 Gauss–Seidel iterations (Figure 2.2 in [2]). The initial vector was $v_1 = [1, \ldots, 1]^T / \sqrt{N}$. Figure 2 shows the rightmost part of the spectrum of A and the spectrum of T_C for $\alpha = 5$. It also shows the radius ρ of the circle centered at 1, enclosing the nonzero eigenvalues of T_C. Figure 1 compares the spectra of T_C and M_C for $\alpha = 5$. Note that ρ is a measure of the separation of the eigenvalues of T_C and M_C. From the results in Table 2, the following conclusions can be drawn.

1. If α is moved away from λ, then the linear systems are easier to solve since $\|G\|_1$ is smaller.

2. If α is moved to the right, then the eigenvalues of T_C are less well separated since ρ is larger.

3. There is an optimum ($\alpha = 5$) that makes the residual norm minimal. For this value, both ρ and $\|G\|$ are small, and this explains the faster convergence for this case. As we shall see in Theorem 4.2, $\|\mathcal{Q}_1 G \mathcal{Q}_1\|$ should be considered to explain the convergence.

TABLE 2
One iteration of the restarted Arnoldi algorithm for the Olmstead model.

α	Residual norms	$\|G\|_1$	$\|\mathcal{Q}_1 G \mathcal{Q}_1\|_1$	ρ
λ_1	$8\,10^{2}$	$2\,10^{18}$	$2\,10^{18}$	0
4	$2\,10^{0}$	153	90	0.65
5	$3\,10^{-5}$	13	17	0.73
6	$3\,10^{-4}$	3	4.8	0.78

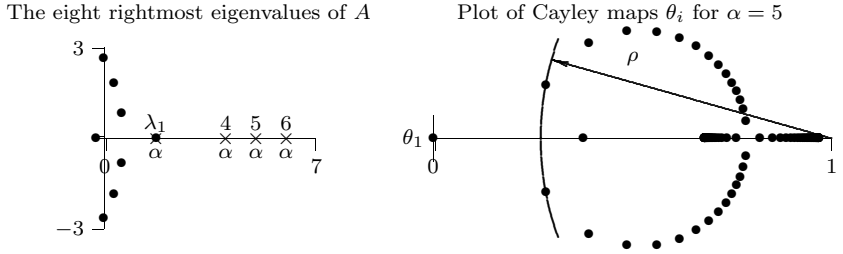


FIG. 2. *Mapping of eigenvalues of the Olmstead model.*

4.3. Asymptotic convergence of the restarted Arnoldi method. Let us now concentrate on the convergence of one iteration of Algorithm 1. The Arnoldi process computes approximate eigenpairs $(\hat{\mu}_i, \hat{x}_i)$ with residual

$$r_i = A\hat{x}_i - \hat{\mu}_i B\hat{x}_i \perp \mathcal{K}_m.$$

In the work by Saad [20], the convergence of an eigenvector x_i in m Arnoldi steps is expressed in terms of

$$\|(I - \mathcal{P}_{\mathcal{K}})x_i\|_2, \tag{26}$$

where $\mathcal{P}_{\mathcal{K}}x_i$ with $\mathcal{P}_{\mathcal{K}} = V_m V_m^H$ denotes the orthogonal projection of x_i on the space $\mathcal{K}_m$. Equation (26) can be viewed as the sine of the acute angle between x_i and the space $\mathcal{K}_m$. In general, the computed eigenvector $\hat{x}_i$ is an approximation of $\mathcal{P}_{\mathcal{K}}x_i$ and the actual error $\|\hat{x}_i - x_i\|_2$ is larger than (26). Using standard perturbation analysis (e.g., [3, Chapters 2 and 4]), one can bound $|\mu_1 - \hat{\mu}_1|$ and $\|x_1 - \hat{x}_1\|_2$ from the residual norm $\|(H_m - \mu_1 F_m)z_1\|_2$ with $z_1 = V_m^H x_1 / \|V_m^H x_1\|_2$. This residual can be bounded as follows.

LEMMA 4.1. *Let (μ, x) be an eigenpair of $Ax = \mu Bx$. Let V_m, H_m, and F_m be computed by one iteration in Algorithm* 1; *then for $z = V_m^H x / \|V_m^H x\|_2$,*

$$\|(H_m - \mu F_m)z\|_2 \leq \gamma \frac{\|(I - \mathcal{P}_{\mathcal{K}})x\|_2}{\|\mathcal{P}_{\mathcal{K}}x\|_2}$$

with γ some constant.

Proof. See [21, p. 130] for the standard case $B = I$. A similar expression is derived for the generalized case by computing an upper bound on $\|(A_m - \mu B_m)\mathcal{P}_{\mathcal{K}}x\|_2$ in equation (4.23) in [21] with $A_m = \mathcal{P}_{\mathcal{K}}A\mathcal{P}_{\mathcal{K}} = VH_mV^H$ and $B_m = \mathcal{P}_{\mathcal{K}}B\mathcal{P}_{\mathcal{K}} = VF_mV^H$ and $\gamma = \|\mathcal{P}_{\mathcal{K}}(A - \mu B)(I - \mathcal{P}_{\mathcal{K}})\|_2$. ∎

This lemma shows that it is meaningful to derive upper bounds to (26) to judge the convergence rate of Arnoldi's process. The following theorem gives such an upper bound when $\mathcal{K}_m$ is a Krylov space of M_{C}.

THEOREM 4.2. *Let $\mathcal{P}_1$ be the spectral projector $\mathcal{P}_1 = x_1 y_1^H$ with $x_1^H y_1 = 1$ and let $\mathcal{Q}_1$ be its spectral complement. Let T_{C} be diagonalizable with Jordan canonical form XZX^{-1}. Consider $v_1 = \delta_x x_1 + \delta_t t$ with $\mathcal{P}_1 t = 0$ and $\|v_1\|_2 = \|x_1\|_2 = \|t\|_2 = 1$. Then for any $\zeta \in \mathbb{C}$, and if $\theta_1 = 0$,*

$$\|x_1 - \mathcal{P}_{\mathcal{K}} x_1\|_2 \leq |\gamma| (\mathrm{cond}_2(X))^{m-1} \left(\max_{j>1} \left| \frac{\zeta - \theta_j}{\zeta - \theta_1} \right| + \|\mathcal{Q}_1 G \mathcal{Q}_1\|_2 \max_{j>1} \left| \frac{\theta_j}{\zeta - \theta_1} \right| \right)^{m-1}, \tag{27}$$

with

$$\gamma = \frac{\delta_t}{\delta_x + \delta_t y_1^H (\zeta I - M_{\mathrm{C}})^{m-1} t / (\zeta - \theta_1)^{m-1}}. \tag{28}$$

Note that for normal matrices, $\mathcal{P}_1$ and $\mathcal{Q}_1$ are orthogonal projectors, since $y_1 = x_1$, and $\mathrm{cond}_2(X) = 1$.

Proof. Consider the vector v_1 as defined above. Following [21, Theorem 3.1], the orthogonal projection $\mathcal{P}_{\mathcal{K}} x_1$ makes $\|x_1 - w\|_2$ minimal for all $w \in \mathcal{K}_m(M_{\mathrm{C}}, v_1)$. Therefore, $\|x_1 - \mathcal{P}_{\mathcal{K}} x_1\|_2 \leq \|x_1 - w\|_2$ for any $w \in \mathcal{K}_m(M_{\mathrm{C}}, v_1)$, so also for

$$w = \frac{\gamma}{\delta_t (\zeta - \theta_1)^{m-1}} (\zeta I - M_{\mathrm{C}})^{m-1} v_1.$$

We shall first prove that

$$\begin{aligned} \|\delta_t (\zeta - \theta_1)^{m-1} / \gamma (x_1 - w)\|_2 \leq |\delta_t| \, & (\|\mathcal{Q}_1 (\zeta I - T_{\mathrm{C}}) \mathcal{Q}_1\|_2 \\ & + \|\mathcal{Q}_1 G \mathcal{Q}_1\|_2 \, \|\mathcal{Q}_1 T_{\mathrm{C}}\|_2)^{m-1}. \end{aligned} \tag{29}$$

From $(\zeta I - M_{\mathrm{C}}) = (\zeta I - T_{\mathrm{C}}) + G T_{\mathrm{C}}$, $\theta_1 = 0$, and $v_1 = \delta_x x_1 + \delta_t t$, we derive that

$$(\zeta I - M_{\mathrm{C}})^{m-1} v_1 = \delta_x (\zeta - \theta_1)^{m-1} x_1 + \delta_t (\zeta I - M_{\mathrm{C}})^{m-1} t. \tag{30}$$

To prove (29), observe that $I = \mathcal{P}_1 + \mathcal{Q}_1$, $\mathcal{Q}_1 \cdot \mathcal{P}_1 = 0$, and $t = \mathcal{Q}_1 t$. Hence

$$(\zeta I - M_{\mathrm{C}})^{m-1} t = \mathcal{P}_1 (\zeta I - M_{\mathrm{C}})^{m-1} t + \mathcal{Q}_1 (\zeta I - M_{\mathrm{C}})^{m-1} \mathcal{Q}_1 t,$$

with

$$\mathcal{P}_1 (\zeta I - M_{\mathrm{C}})^{m-1} t = (y_1^H (\zeta I - M_{\mathrm{C}})^{m-1} t) x_1$$

and

$$\begin{aligned} \mathcal{Q}_1 (\zeta I - M_{\mathrm{C}})^{m-1} \mathcal{Q}_1 t &= (\mathcal{Q}_1 (\zeta - M_{\mathrm{C}}) \mathcal{Q}_1)^{m-1} t \\ &= (\mathcal{Q}_1 (\zeta I - T_{\mathrm{C}}) \mathcal{Q}_1 + \mathcal{Q}_1 G T_{\mathrm{C}} \mathcal{Q}_1)^{m-1} t. \end{aligned}$$

Thus (30) becomes

$$\begin{aligned} (\zeta I - M_{\mathrm{C}})^{m-1} v_1 = & (\delta_x (\zeta - \theta_1)^{m-1} + \delta_t y_1^H (\zeta I - M_{\mathrm{C}})^{m-1} t) x_1 \\ & + \delta_t (\mathcal{Q}_1 (\zeta I - T_{\mathrm{C}}) \mathcal{Q}_1 + \mathcal{Q}_1 G T_{\mathrm{C}} \mathcal{Q}_1)^{m-1} t, \end{aligned}$$

from which (29) follows. Since $\|\mathcal{Q}_1 T_{\mathrm{C}}\|_2 \leq \mathrm{cond}_2(X) \max_{j>1} |\theta_j|$ and $\|\mathcal{Q}_1 (\zeta I - T_{\mathrm{C}})\|_2 \leq \mathrm{cond}_2(X) \max_{j>1} |\zeta - \theta_j|$, (27) immediately follows from (29). ∎

The theorem covers two extremal cases.

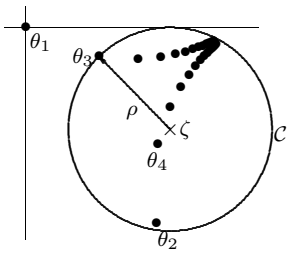


FIG. 3. *Illustration of the choice of* ζ.

1. If $G = 0$, then $\gamma = \delta_t/\delta_x$ since $M_{\mathrm{C}} = T_{\mathrm{C}}$ and $y_1^H(\zeta I - T_{\mathrm{C}})t = 0$. In this case,

$$\|x_1 - \mathcal{P}_{\mathcal{K}} x_1\| \leq \mathrm{cond}_2(X)^{m-1} \left|\frac{\delta_t}{\delta_x}\right| \max_{j>1} \left|\frac{\zeta - \theta_j}{\zeta - \theta_1}\right|^{m-1},$$

which depends only on spectral properties of T_{C}, just as in the shift-invert Arnoldi method. $\|x_1 - \mathcal{P}_{\mathcal{K}} x_1\|$ is small if $|\zeta - \theta_j| \ll |\zeta - \theta_1|$ for $j = 2, \ldots, N$. This is equivalent to the condition that there is a circle $\mathcal{C}$ with center ζ that encloses the eigenvalues θ_j, $j = 2, \ldots, N$ of T_{C} without enclosing θ_1; see Figure 3. In fact, $\max_{i>1} |\zeta - \theta_i|$ is the radius ρ of $\mathcal{C}$ such that the ratio $\rho/|\zeta - \theta_1|$ determines the convergence rate. Clearly, the ratio $\rho/|\zeta - \theta_1|$ is smaller when α lies closer to μ_1 (see, e.g., Example 3 for $\zeta = 1$). This explains why one traditionally picks α as close as possible to the eigenvalue μ_1 in shift-invert Arnoldi.

2. If $\|G\| \gg \|T_{\mathrm{C}}\|$ and δ_x is small, then $\zeta I - M_{\mathrm{C}} \simeq G T_{\mathrm{C}}$ and

$$\|x_1 - \mathcal{P}_{\mathcal{K}} x_1\| \leq \mathrm{cond}(X)^{m-1} \frac{(\|\mathcal{Q}_1 G \mathcal{Q}_1\| \max_{j>1} |\theta_j|)^{m-1}}{|y_1^H (G T_{\mathrm{C}})^{m-1} t|}.$$

Thus, roughly speaking, the convergence depends on the ratio $\|\mathcal{Q}_1 G \mathcal{Q}_1\| / \|\mathcal{P}_1 G \mathcal{Q}_1\|$.

Theorem 4.2 can also be used to explain the asymptotic convergence of the restarted Arnoldi method. Assuming that after a number of iterations the vector $\hat{x}_1$ is rich in x_1 such that in the next iteration $\delta_x \gg \delta_t$ and $\lambda \simeq \mu_1$, the error reduction is given by (27) with $\gamma \simeq \delta_t/\delta_x$. Fast asymptotic convergence is thus reached by a compromise between a good separation of eigenvalues of T_{C} (i.e., small ρ) and small $\|\mathcal{Q}_1 G \mathcal{Q}_1\|$. On the one hand, α should lie close to μ_1 to make $\rho/|\zeta - \theta_1|$ small, but on the other $\|\mathcal{Q}_1 G \mathcal{Q}_1\|$ must be small, which often can be obtained by choosing α far away from μ_1. These points agree with the results in Example 3. Shifting α to the right also provides good mapping properties between the eigenvectors x_i of T_{C} and u_k of M_{C}, as was explained at the end of section 3. Note that Theorem 4.2 does not provide a sharp bound, but rather shows the factors that play a role in the convergence.

Remark 4.1. Note that a similar bound on the asymptotic convergence can be derived instead of (27); namely,

$$\|x_1 - \mathcal{P}_{\mathcal{K}} x_1\|_2 \leq |\gamma| \mathrm{cond}_2(X) \left(\max_{j>1} \left|\frac{\zeta - \theta_j}{\zeta - \theta_1}\right| + \mathrm{cond}_2(X) \|\mathcal{Q}_1 G \mathcal{Q}_1\|_2 \max_{j>1} \left|\frac{\theta_j}{\zeta - \theta_1}\right| \right)^{m-1}.$$

This upper bound is more sensitive to the solution of the linear system, but less to the separation of the eigenvalues of T_{C} if $\mathrm{cond}_2(X)$ is large.

4.4. The restarted Davidson method. For comparison, we give a short discussion of Davidson's method. In this method (inner loop in Algorithm 2), the parameter λ is updated before each matrix–vector multiplication $w_i = M_C\hat{x}_1$. In fact, in each step of the method the eigenvector $\hat{x}_1$ becomes richer in the eigenvector x_1. Note that the subspace $\mathcal{K}_m$ spanned by $v_1, \ldots, v_m$ is no longer a Krylov space since the matrix M_C changes in each step. Also note that w_i is computed from the approximate rightmost eigenvector $\hat{x}_1$ and not from the basis vector v_i. A detailed analysis of Davidson's method is given in [14, 13, 24, 4]. Note that if $\lambda = \mu_1$, Theorem 4.2 also holds for the asymptotic convergence of the restarted Davidson method, since in this case Arnoldi and Davidson produce the same space. Nevertheless, we shall see in the examples in section 5 that the choice of α seems to have less influence on the convergence than in the restarted Arnoldi method.

ALGORITHM 2 (a restarted Davidson algorithm applied to M_C).
Given α and $\hat{x}_1$.
repeat
  Set up the linear system solver applied to (4).
  Let $v_1 = \hat{x}_1/\|\hat{x}_1\|$.
  for $i = 1$ **to** m **do**
    Compute $H_i = V_i^H A V_i$ and $F_i = V_i^H A V_i \in \mathbb{C}^{i\times i}$ with $V_i = [v_1, \ldots, v_i]$.
    Compute the eigenpairs $(\hat{\mu}_j, \hat{z}_j)$ of $H_i = \mu F_i z$ by the QZ method.
    Form $\hat{x}_1 = V_i\hat{z}_1$.
    Set $\lambda = \hat{\mu}_1$.
    if $i = m$ **then exit loop**
    Form $w_i = M_C\hat{x}_1$.
    Orthonormalize w_i against $v_1, \ldots, v_i$.
    Let $v_{i+1} = w_i$.
  end for
  Reset α.
until $\|A\hat{x}_1 - \hat{\mu}_1 B\hat{x}_1\|_2 < \epsilon\|\hat{x}_1\|_2$

4.5. How to manage complex arithmetic. In general, μ_1 is complex. Therefore, λ and α can take complex values. This makes complex vectors in Algorithms 1 and 2 inevitable. We now explain how complex arithmetic can be restricted.

The use of shift-invert with complex shift for real matrices has been studied by Parlett and Saad [17] and Ruhe [19]. Following the former approach, the eigenpairs are computed from the Krylov space $\mathcal{K}_m(\mathrm{Re}(M_C), v_1)$ rather than $\mathcal{K}_m(M_C, v_1)$, which involves only real vectors. However, this approach is useful only when $\mathrm{Im}(\lambda)$ is very small, since when α is real,

$$\mathrm{Re}(M_C) = (M_A - \alpha M_B)^{-1}(A - \mathrm{Re}(\lambda)B),$$

and eigenvectors of $Ax = \mu Bx$ corresponding to the eigenvalues close to λ do not correspond well to those of $\mathrm{Re}(M_C)$. On the other hand, the approach by Ruhe is very useful in this application. He uses the subspace $\mathcal{R}$ spanned by

$$\mathcal{B}_{\mathcal{R}} = \{\mathrm{Re}(v_1), \mathrm{Im}(v_1), \ldots, \mathrm{Re}(v_m), \mathrm{Im}(v_m)\},$$

where $v_1, \ldots, v_m$ are the (complex) Arnoldi vectors that span $\mathcal{K}_m(M_C, v_1)$. Because $\mathcal{B}_{\mathcal{R}}$ is real, we refer to projection on $\mathcal{R}$ as *real projection* and projection on $\mathcal{K}_m$ as *complex projection*. In general, the dimension of $\mathcal{R}$ lies between m and $2m$.

One advantage of using $\mathcal{R}$ instead of $\mathcal{K}_m$ is that $V^T AV$ can be computed by real arithmetic. The most important advantage is that the subspace $\mathcal{R}$ contains $\mathcal{K}_m$.

Hence eigenvectors are approximated better by orthogonal projection on $\mathcal{R}$ than on $\mathcal{K}_m$. We will illustrate the difference between real and complex projection in the examples in section 5.

A similar treatment for avoiding complex arithmetic can be used in Davidson's method by splitting the vectors v_i, $i = 1, \ldots, m$ in Algorithm 2 into their real and imaginary parts and adding these parts separately to the subspace. This is suggested by Morgan [13] and Sadkane [24].

Remark 4.2. In each outer iteration of Arnoldi's method and in each step of Davidson's method, λ is set to the rightmost approximate eigenvalue $\hat{\mu}_1$. With this choice, we hope that λ lies close to μ_1 in order to achieve a good link between the eigenvectors of M_{C} and $Ax = \mu Bx$. Suppose that the rightmost eigenvalues of $Ax = \mu Bx$ are complex and $\mu_2 = \bar{\mu}_1$. With complex projection, $\hat{\mu}_1$ and $\hat{\mu}_2$ are the rightmost eigenvalues of $H_m z = \mu F_m z$. Because H_m and F_m are complex, $\hat{\mu}_2 = \bar{\hat{\mu}}_1$ is not true in general. Both $\hat{\mu}_1$ and $\hat{\mu}_2$ can be selected as the next λ. We always choose λ as the computed rightmost eigenvalue in the upper half plane. In the case of real projection, this selection strategy is not necessary since H_m is real and so $\hat{\mu}_1 = \bar{\hat{\mu}}_2$.

5. Numerical examples. For the restarted Arnoldi method, we illustrate the influence of m on the convergence rate (Example 4). We also compare the restarted Arnoldi and Davidson methods (Examples 5 and 6) and real and complex projection (Example 5). In our experiments, the first iteration of Algorithms 1 and 2 consists of m steps of Arnoldi's process applied to M_{C} with $\alpha = \lambda = 0$ and with $v_1 = [1, \ldots, 1]^T/\sqrt{N}$.

Example 4. This example originates from the system

$$\begin{cases} u_t = u_{ss} + 5v_{ss}, \\ v_t = v_{ss} + u \end{cases}$$

with spatial coordinate $s \in [0, 1]$ subject to homogeneous Dirichlet boundary conditions. After discretization with central differences with grid size $h = \frac{1}{N/2}$, the equations are written as $\dot{x} = f(x)$ with $x = [u_1, v_1, u_2, v_2, \ldots, u_{N/2}, v_{N/2}]^T$ and the problem leads to a standard eigenvalue problem. The size of the Jacobian matrix is $N = 3070$ and $\mu_{1,2} \simeq -9.87 \pm 7.02i$. We used Arnoldi's method (Algorithm 1) with complex projection. The linear systems were solved by one Gauss–Seidel multigrid V-cycle without presmoothing and with two postsmoothing steps. Formula (25) was used to compute $w = M_{\mathrm{C}} v$. We examine the choice of m in the restarted Arnoldi process. In each iteration of Algorithm 1, $\alpha = 0$ was used, so α was not reset. The history of the residual norm $r_1 = \|A\hat{x}_1 - \lambda\hat{x}_1\|_2/\|\hat{x}_1\|_2$ and the execution times for one processor of the IBM SP2 are given in Table 3 for various values of m.

The following observations are now made. The asymptotic convergence is almost linear and the convergence speed increases with increasing m, which was predicted by Theorem 4.2. Nevertheless, the total execution time increases for larger m. This can be explained by the fact that at least two iterations are necessary to find a λ that lies in the neighborhood of μ_1, independent of m, which is very expensive for large m. In fact, a large number of inner Arnoldi steps find the eigenvectors of M_{C} better, but do not necessarily find a better λ, since the eigenvectors of M_{C} do not correspond very well to those of the original eigenvalue problem in the first iterations.

Example 5. Here we compare Arnoldi's and Davidson's methods (Algorithms 1 and 2) for different strategies for choosing α, and we also compare real and complex projection. We solved the problem in Example 4 for $N = 3070$ with the same multigrid solver. In a first run, α was reset to λ in each (outer) iteration and in a second run

TABLE 3
Influence of m on the convergence of the restarted Arnoldi method: residual norms and execution times for Example 4.

m	3	5	10	20	30
iteration 1	$3\,10^{5}$	$1\,10^{5}$	$1\,10^{4}$	$7\,10^{3}$	$6\,10^{3}$
2	$2\,10^{2}$	$2\,10^{1}$	$9\,10^{0}$	$6\,10^{0}$	$5\,10^{0}$
3	$7\,10^{3}$	$1\,10^{0}$	$9\,10^{-3}$	$3\,10^{-3}$	$2\,10^{-3}$
4	$4\,10^{2}$	$2\,10^{-2}$	$9\,10^{-7}$	$3\,10^{-7}$	$8\,10^{-8}$
5	$4\,10^{2}$	$4\,10^{-4}$	$1\,10^{-9}$	$1\,10^{-9}$	$1\,10^{-9}$
6	$3\,10^{1}$	$8\,10^{-6}$			
7	$5\,10^{0}$	$2\,10^{-7}$			
8	$6\,10^{-1}$	$4\,10^{-9}$			
9	$9\,10^{-2}$				
10	$7\,10^{-3}$				
11	$8\,10^{-4}$				
12	$2\,10^{-4}$				
13	$8\,10^{-5}$				
14	$2\,10^{-5}$				
15	$2\,10^{-6}$				
16	$3\,10^{-7}$				
17	$1\,10^{-7}$				
18	$2\,10^{-8}$				
19	$2\,10^{-9}$				
time (sec)	2.4	2.2	3.8	12	25

TABLE 4
Comparison of the restarted Arnoldi and Davidson methods: results for Example 5.

	$\alpha = \lambda$				$\alpha = 0$			
	Arnoldi		Davidson		Arnoldi		Davidson	
projection	complex	real	complex	real	complex	real	complex	real
iter. 1	$1.4\,10^{4}$	$1.4\,10^{4}$	$1.4\,10^{4}$	$1.4\,10^{4}$	$1.4\,10^{4}$	$1.4\,10^{4}$	$1.4\,10^{4}$	$1.4\,10^{4}$
2	$5.4\,10^{2}$	$5.4\,10^{2}$	$3.0\,10^{2}$	$5.1\,10^{2}$	$8.6\,10^{0}$	$8.6\,10^{0}$	$6.4\,10^{-2}$	$6.8\,10^{-3}$
3	$7.7\,10^{1}$	$1.9\,10^{1}$	$1.4\,10^{1}$	$3.0\,10^{-3}$	$8.7\,10^{-3}$	$3.6\,10^{-4}$	$1.3\,10^{-7}$	$6.1\,10^{-9}$
4	$6.9\,10^{-1}$	$4.6\,10^{-2}$	$3.7\,10^{-6}$	$2.1\,10^{-9}$	$9.2\,10^{-7}$	$1.8\,10^{-9}$	$1.2\,10^{-9}$	
5	$2.7\,10^{-2}$	$2.9\,10^{-8}$	$1.3\,10^{-9}$		$1.2\,10^{-9}$			
6	$1.1\,10^{-3}$	$1.6\,10^{-9}$						
7	$7.8\,10^{-5}$							
8	$1.5\,10^{-6}$							
9	$5.3\,10^{-7}$							
10	$1.8\,10^{-7}$							
11	$3.5\,10^{-8}$							
12	$8.1\,10^{-9}$							
iterations	12	5	5	5	5	4	4	3
time (sec.)	10	5.3	4.9	4.4	3.8	3.4	3.8	3.4
$\rho(G)$	3.5				0.11			

α was kept equal to zero. The subspace dimension m is set to 10. Table 4 shows the residual norm r_1 per outer iteration. Also, the number of iterations, the execution times, and the spectral radius of G in the last iteration are given.

It appears that in the first iteration(s), real and complex projection have the same residuals. This is because λ takes real values in the first iterations, and so there is no difference between real and complex projection.

Recall from the numerical results from Example 4 that frequent updates of λ often lead to faster convergence. This explains why Davidson's method converges faster than Arnoldi's method. However, note that one iteration of Davidson's method

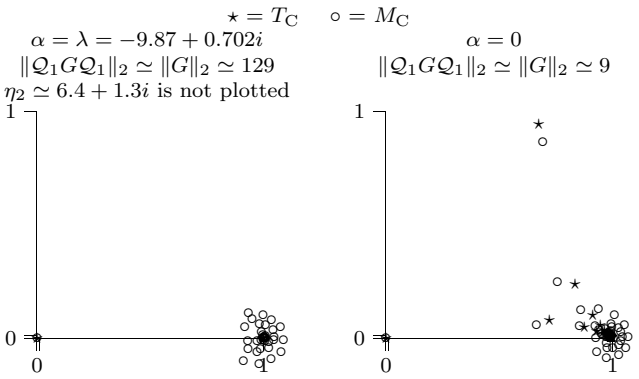


FIG. 4. *Influence of the choice of α on the convergence of the restarted Arnoldi method: spectra of T_C and M_C and norms of G for Example 5 with $\lambda = \mu_1$ and $N = 46$.*

is more expensive because of the orthogonal projections in the inner iterations of the algorithm.

In Arnoldi's method, real projection produces smaller residuals than complex projection for the same number of iterations, especially when α lies close to λ. In Davidson's method, the advantage of real projection is less pronounced.

This example also illustrates the influence of moving α to the right of the spectrum. To explain the difference in convergence speed for Arnoldi's method (12 iterations for $\alpha = \lambda$ and 6 iterations for $\alpha = 0$), we refer to Theorem 4.2 and section 3. Figure 4 shows the spectra of T_C and M_C for $N = 46$, $\lambda = \mu_1$, and $\alpha = \lambda$ and $\alpha = 0$. Note that $\alpha = 0$ lies further from the spectrum than $\alpha = \lambda$. First consider $\alpha = \lambda = \mu_1$. Recall from Remark 3.2 that $T_C = \mathcal{Q}_1$ when $\alpha = \lambda$. Hence, T_C has eigenvalues zero and 1. The large eigenvalue $\eta_2 = 6.4 + 1.3i$ seems to hinder the convergence. $\rho(G)$ and $\|\mathcal{Q}_1 G \mathcal{Q}_1\|_2$ are quite large. For $\alpha = 0$, the spectral radius $\rho(G)$ and $\|\mathcal{Q}_1 G \mathcal{Q}_1\|$ are rather small. This makes the second term in (27) smaller. Although moving α away from λ makes the eigenvalues of T_C less separated, they are separated enough to make the first term in (27) small for a well-chosen ζ. A similar conclusion holds for Davidson's method, but the choice of α seems to be less important.

Example 6. We have studied the influence of the choice of α on the convergence of the restarted Arnoldi method and restarted Davidson method for the Olmstead eigenvalue problem from Example 3 with rightmost eigenvalue $\mu_1 = 1.6905$ (see Figure 2). The linear systems were solved approximately by 10 Gauss–Seidel iterations. Formula (24) was used to compute $w = M_C v$. The algorithms were run with $m = 10$. Since λ is real, it is sufficient to compare the results for real projection. It is clear from Table 5 that Davidson is less susceptible to α, while Arnoldi's method is very sensitive to α. We found that both methods stagnate for $\sigma = \hat{\lambda} + 2$.

6. Conclusions. This paper presents and analyzes the use of $M_C = (M_A - \alpha M_B)^{-1}(A - \lambda B)$ in the restarted Arnoldi method. The analysis is general in the sense that no restrictions are imposed on the linear system solver that is used to solve $(A - \alpha B)z = b$, and is particularly useful to explain the asymptotic convergence of restarted Arnoldi and Davidson. If the eigenvalues are complex λ is also complex, and this fact involves complex arithmetic. However, the complex work is reduced by using a real α and by the use of real projection. If a good iterative linear system solver is used, then Arnoldi's method quickly converges to a good approximation of

TABLE 5
Influence of the choice of α on the convergence of the restarted Arnoldi and Davidson methods: results for the Olmstead model in Example 6.

	$\alpha = \mathrm{Re}(\lambda) + 2.5$		$\alpha = \mathrm{Re}(\lambda) + 3$		$\alpha = \mathrm{Re}(\lambda) + 4$	
iteration	Arnoldi	Davidson	Arnoldi	Davidson	Arnoldi	Davidson
1	$2.1\,10^{2}$	$2.1\,10^{2}$	$2.1\,10^{2}$	$2.1\,10^{2}$	$2.1\,10^{2}$	$2.1\,10^{2}$
2	$3.0\,10^{-1}$	$5.7\,10^{-1}$	$3.1\,10^{-1}$	$7.3\,10^{-1}$	$4.7\,10^{-1}$	$1.4\,10^{0}$
3	$3.4\,10^{-1}$	$3.0\,10^{-1}$	$6.8\,10^{-2}$	$1.3\,10^{-2}$	$4.3\,10^{-3}$	$5.2\,10^{-2}$
4	$1.0\,10^{-1}$	$6.1\,10^{-2}$	$6.3\,10^{-4}$	$2.1\,10^{-5}$	$3.8\,10^{-5}$	$4.0\,10^{-4}$
5	$5.0\,10^{-2}$	$1.4\,10^{-3}$	$2.7\,10^{-5}$	$4.3\,10^{-8}$	$3.3\,10^{-7}$	$2.4\,10^{-6}$
6	$2.5\,10^{-2}$	$1.1\,10^{-4}$	$5.5\,10^{-6}$	$4.6\,10^{-11}$	$3.9\,10^{-9}$	$1.0\,10^{-8}$
7	$5.2\,10^{-3}$	$1.3\,10^{-5}$	$3.5\,10^{-9}$			$7.9\,10^{-11}$
8	$2.3\,10^{-3}$	$5.5\,10^{-7}$				
9	$1.3\,10^{-4}$	$7.4\,10^{-8}$				
10	$2.4\,10^{-5}$	$6.0\,10^{-9}$				
11	$1.4\,10^{-6}$					
12	$1.1\,10^{-7}$					
13	$4.7\,10^{-9}$					
iterations	13	10	7	6	6	7
time (sec.)	0.15	0.21	0.09	0.14	0.08	0.15
$\|\mathcal{Q}_1 G \mathcal{Q}_1\|_1$	63		27		6.7	

μ_1. Davidson's method quickly converges to accurate estimates of μ_1, in general faster than Arnoldi's method. In our applications, however, Arnoldi's method is competitive with Davidson's method.

There is no general guideline for choosing α. When a direct linear system solver is used all arguments are in favor of choosing α close to μ_1, but from the experiments it looks as though an α further away gives smoother convergence behavior and shorter overall solution time with an iterative solver. Of course, it depends on the system solver whether and how far α should be moved away. Perhaps this question needs further investigation. The convergence of the generalized problem depends on the matrix B and thus it is much harder to formulate a general convergence result in this case.

Acknowledgments. This paper presents research results from the Belgian Programme on Interuniversity Poles of Attraction (IUAP 17), initiated by the Belgian State–Prime Minister's Service–Federal Office for Scientific, Technical and Cultural Affairs. The scientific responsibility rests with its authors.

The authors are grateful to the referees for their careful reading and the many comments that improved the quality of the paper.

REFERENCES

[1] Z. Bai, *A spectral transformation block Lanczos algorithm for solving sparse non-Hermitian eigenproblems*, in Proc. of the Fifth SIAM Conference on Applied Linear Algebra, Philadelphia, PA, J. Lewis, ed., 1994, SIAM, pp. 307–311.

[2] R. Barrett, M. Berry, T. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. Van der Vorst, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM, Philadelphia, PA, 1994.

[3] F. Chatelin, *Eigenvalues of Matrices*, John Wiley, New York, 1993.

[4] M. Crouzeix, B. Philippe, and M. Sadkane, *The Davidson method*, SIAM J. Sci. Comput., 15 (1994), pp. 62–76.

[5] J. Cullum, W. Kerner, and R. Willoughby, *A generalised nonsymmetric Lanczos procedure*, Comput. Phys. Comm., 53 (1989), pp. 19–48.

[6] T. Ericsson and A. Ruhe, *The spectral transformation Lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems*, Math. Comp., 35 (1980), pp. 1251–1268.
[7] R. Freund and N. Nachtigal, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.
[8] G. Golub and C. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[9] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Nat. Bur. Stand., 45 (1950), pp. 255–282.
[10] K. Meerbergen, A. Spence, and D. Roose, *Shift-invert and Cayley transforms for detection of rightmost eigenvalues of nonsymmetric matrices*, BIT, 34 (1994), pp. 409–423.
[11] H. Mittelmann, K.-T. Chang, D. Jankowski, and G. Neitzel, *Iterative solution of the eigenvalue problem in Hopf bifurcation for the Boussinesq equations*, SIAM J. Sci. Comput., 15 (1994), pp. 704–712.
[12] H. Mittelmann, C. Law, D. Jankowski, and P. Neitzel, *A large, sparse, and indefinite generalized eigenvalue problem from fluid mechanics*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 411–424.
[13] R. Morgan, *Generalisations of Davidson's method for computing eigenvalues of large nonsymmetric matrices*, J. Comput. Phys., 101 (1992), pp. 287–291.
[14] R. Morgan and D. Scott, *Generalizations of Davidson's method for computing eigenvalues of sparse symmetric matrices*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 817–825.
[15] R. Morgan and D. Scott, *Preconditioning the Lanczos algorithm for sparse symmetric eigenvalue problems*, SIAM J. Sci. Comput., 14 (1993), pp. 585–593.
[16] W. Olmstead, W. Davis, S. Rosenblat, and W. Kath, *Bifurcation with memory*, SIAM J. Appl. Math., 40 (1986), pp. 171–188.
[17] B. Parlett and Y. Saad, *Complex shift and invert strategies for real matrices*, Linear Algebra Appl., 88/89 (1987), pp. 575–595.
[18] A. Ruhe, *Rational Krylov sequence methods for eigenvalue computation*, Linear Algebra Appl., 58 (1984), pp. 391–405.
[19] A. Ruhe, *The rational Krylov algorithm for nonsymmetric eigenvalue problems,* III: *Complex shifts for real matrices*, BIT, 34 (1994), pp. 165–176.
[20] Y. Saad, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebra Appl., 34 (1980), pp. 269–295.
[21] Y. Saad, *Numerical methods for large eigenvalue problems*, in Algorithms and Architectures for Advanced Scientific Computing, Manchester University Press, Manchester, U.K., 1992.
[22] Y. Saad, *SPARSKIT: A Basic Tool Kit for Sparse Matrix Computations*, Tech. report 90-20, Research Institute for Advanced Computer Science, NASA Ames Research Center, Moffet Field, CA, 1990.
[23] Y. Saad and M. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[24] M. Sadkane, *Block-Arnoldi and Davidson methods for unsymmetric large eigenvalue problems*, Numer. Math., 64 (1993), pp. 195–211.
[25] R. Silverman, *Introductory Complex Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1967.
[26] G. Sleijpen, D. Fokkema, and H. Van Der Vorst, *Bi-CGSTAB (ℓ) and other hybrid* Bi-CG *methods*, Numer. Algorithms, 7 (1994), pp. 75–109.
[27] G. Sleijpen and H. Van der Vorst, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1995), pp. 401–425.
[28] J. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, U.K., 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 21–29, January 1997

002

RELATIVE RESIDUAL BOUNDS FOR THE EIGENVALUES OF A HERMITIAN SEMIDEFINITE MATRIX*

ZLATKO DRMAČ† AND VJERAN HARI‡

Dedicated to the memory of Branko Najman (1949–1996).

Abstract. Let H be a Hermitian matrix, X an orthonormal matrix, and $M = X^*HX$. Then the eigenvalues of M approximate some eigenvalues of H with an absolute error bounded by $\|R\|_2$, $R = HX - XM$. This work contains estimates of $|\lambda-\mu|/|\mu|$ and $|\lambda-\mu|/|\lambda|$, where μ, λ is a matching pair of the eigenvalues of M and H when H is semidefinite. The general bound is expressed in terms of sines of the canonical angles between certain subspaces associated with H and X. A more refined quadratic bound which uses the relative distances between eigenvalues is also proved.

Key words. residual bounds, relative error, eigenvalue location

AMS subject classifications. 65F15, 65G05

PII. S0895479895284002

Introduction. Let $H \in \mathbf{C}^{n\times n}$ be Hermitian, $X \in \mathbf{C}^{n\times m}$ orthonormal, and

$$M = X^*HX, \quad R = HX - XM.$$

Let $\mathcal{X} = \mathcal{R}(X)$ be the range of X and $\sigma(H) = \{\lambda_j\}$, $\sigma(M) = \{\mu_k\}$ the spectra of H, M, respectively. For the eigenvalues of H and M we assume

$$\lambda_1 \geq \cdots \geq \lambda_n \quad \text{and} \quad \mu_1 \geq \cdots \geq \mu_m,$$

respectively. If $\mathcal{X}$ is an invariant subspace of H, then $R = 0$ and $\sigma(M) \subseteq \sigma(H)$. If $\mathcal{X}$ is close to an invariant subspace of H, then R is close to 0 and each eigenvalue of M is close to an eigenvalue of H. This claim is quantified by the following classical results.

THEOREM 0.1 (see Kahan [8]). *There are eigenvalues λ_{j_k}, $k = 1, \ldots, m$ of H such that*

$$| \lambda_{j_k} - \mu_k | \leq \|R\|_2, \quad k = 1, \ldots, m. \tag{1}$$

THEOREM 0.2 (see Sun [13]). *Let $\mathcal{Y} = \mathcal{R}(Y)$ be an invariant subspace of H with orthonormal basis $Y \in \mathbf{C}^{n\times m}$. Let $\lambda_{j_1} \geq \cdots \geq \lambda_{j_m}$ be the eigenvalues of Y^*HY, and $\Lambda_{\mathcal{Y}} = \operatorname{diag}(\lambda_{j_1}, \ldots, \lambda_{j_m})$, $\Lambda_{\mathcal{X}} = \operatorname{diag}(\mu_1, \ldots, \mu_m)$. If for some $\alpha, \beta \in \mathbf{R}$ and $\delta_0 > 0$,*

$$\sigma(M) \subset [\alpha, \beta], \quad \sigma(H) \setminus \sigma(Y^*HY) \subset (-\infty, \alpha - \delta_0] \bigcup [\beta + \delta_0, +\infty) \quad \textit{(or vice versa)}$$

and if $\rho \equiv \|R\|_2/\delta_0 < 1$, then for any unitarily invariant norm $\|\cdot\|$,

$$\|\Lambda_{\mathcal{Y}} - \Lambda_{\mathcal{X}}\| \leq \frac{1}{\sqrt{1-\rho^2}} \frac{\|R\|_2 \|R\|}{\delta_0}.$$

* Received by the editors April 3, 1995; accepted for publication (in revised form) by N. J. Higham December 12, 1995.
http://www.siam.org/journals/simax/18-1/28400.html
† Department of Computer Science, University of Colorado, Boulder, CO 80309-0430 (zlatko@cs.colorado.edu). The research of this author was supported by National Science Foundation grant ASC-9357812 and Department of Energy grant DE-FG03-94ER25215.
‡ Department of Mathematics, University of Zagreb, Bijenička 30, 41000 Zagreb, Croatia (hari@math.hr).

Furthermore, if $\delta \equiv \min\{|\mu-\lambda| \ : \ \mu \in \sigma(M),\ \lambda \in \sigma(H) \setminus \sigma(Y^*HY)\} > 0$, *and* $\rho_F \equiv \|R\|_F/\delta < 1$, *then*

$$\|\Lambda_{\mathcal{Y}} - \Lambda_{\mathcal{X}}\|_F \leq \frac{1}{\sqrt{1-\rho_F^2}} \frac{\|R\|_F^2}{\delta}. \tag{2}$$

In the above estimates $\|\cdot\|_2$ and $\|\cdot\|_F$ denote the spectral and the Frobenius matrix norms, respectively. Theorem 0.1 is a corollary of a more general result [8] that treats the case of an arbitrary $m \times m$ Hermitian matrix in place of M and an arbitrary $n \times m$ full column rank matrix instead of X. For other bounds of this kind, see [10], [3], [12], [11], [2]. All these results estimate absolute distance between two matching eigenvalues of M and H. The bounds use the norm of the residual (thus, an absolute measure of R) and the (absolute) distance δ between $\sigma(M)$ and the "nonmatching" part of $\sigma(H)$. Often the assumptions of Theorem 0.2 are not met (see Example 2.4 below). Note also that the bound in (2) becomes large (and thus useless) if δ gets tiny enough.

Our aim is to bound $|\lambda_{j_k} - \mu_k|/|\mu_k|$ and consequently $|\lambda_{j_k} - \mu_k|/|\lambda_{j_k}|$, where μ_k , λ_{j_k}, $k = 1, \ldots, m$ are appropriately matching pairs of eigenvalues. Note that the appropriate relative error bounds derived from Theorem 0.1 and Theorem 0.2 can become useless for small or close eigenvalues.

We seek bounds dependent on relative quantities such as (canonical) angles between subspaces and relative distances between eigenvalues. The first such estimates were proved in [6] and the following two theorems summarize them.

THEOREM 0.3. *Let* H *be nonsingular and* $\mathcal{Y} = H\mathcal{X}$, $\mathcal{Z} = H^{-1}\mathcal{X}$. *There are at least* m *eigenvalues* λ_{j_k}, $k = 1, \ldots, m$ *of* H *for which*

$$\frac{|\lambda_{j_k} - \mu_k|}{|\lambda_{j_k}|} \leq \|(I - P_{\mathcal{X}})P_{\mathcal{Y},\mathcal{Z}} + P_{\mathcal{X}}(I - P_{\mathcal{Y},\mathcal{Z}})\|_2, \quad k = 1, \ldots, m \tag{3}$$

holds, provided that the right-hand side is less than one. Here $P_{\mathcal{X}}$ *denotes the orthogonal projector on* $\mathcal{X}$ *and* $P_{\mathcal{Y},\mathcal{Z}}$ *is the projector on* $\mathcal{Y}$ *along* $\mathcal{Z}^{\perp}$. *Furthermore, the right-hand side in* (3) *is bounded by* $\sin\xi + \tan\zeta$, *where* ξ *is the maximal acute angle between* $\mathcal{X}$ *and* $\mathcal{Y}$ *and* ζ *is the maximal acute angle between* $\mathcal{Y}$ *and* $\mathcal{Z}$.

In Theorem 0.3 $\mathcal{Z}^{\perp}$ denotes the orthogonal complement of $\mathcal{Z}$. By an abuse of notation we have denoted by H the operator whose matrix in the standard basis is H. This will be repeated with the matrix L below.

THEOREM 0.4. *Let* $H = LL^*$ *be positive definite,* $\mathcal{Y}_L = L^*\mathcal{X}$, $\mathcal{Z}_L = L^{-1}\mathcal{X}$ *and let* ψ *be the maximal acute angle between* $\mathcal{Y}_L$ *and* $\mathcal{Z}_L$. *Then there are at least* m *eigenvalues* λ_{j_k}, $k = 1, \ldots, m$ *of* H *for which*

$$\frac{|\lambda_{j_k} - \mu_k|}{\lambda_{j_k}} \leq \frac{\sin\psi}{1 - \sin\psi}, \quad k = 1, \ldots, m \tag{4}$$

holds, provided that $\sin\psi/(1-\sin\psi)$ *is less than one.*

If $\mathcal{X}$ is invariant for H from Theorem 0.4, we have $LL^*\mathcal{X} \subseteq \mathcal{X}$, that is, $\mathcal{Y}_L \subseteq \mathcal{Z}_L$. Since $\mathcal{Y}_L$ and $\mathcal{Z}_L$ have the same dimension, we have $\mathcal{Y}_L = \mathcal{Z}_L$ and $\psi = 0$. Note that the converse is also true, hence we have $\psi = 0$ iff $\mu_k = \lambda_{j_k}$, $k = 1, \ldots, m$. The bounds (3), (4) are derived using the backward perturbation $\delta H = RX^* + XR^*$, relative eigenvalue perturbation estimates from [4], [14], [5], and special geometric structure of the operators $\delta H H^{-1}$ and $L^{-1}\delta H L^{-*}$.

In this paper we derive new bounds for $|\lambda_{j_k} - \mu_k|/|\mu_k|$ provided the matrix H is semidefinite. One of them is similar to that of Theorem 0.4 but uses another pair of subspaces and a slightly different angle function. The others assume sufficiently small canonical angles between subspaces but in return deliver bounds which are quadratic in the sine of the maximal angle between subspaces.

1. Relative estimates for semidefinite matrices. Here we derive a new relative a posteriori bound for the eigenvalues of a Hermitian semidefinite matrix H obtained from the subspace $\mathcal{X} \subset \mathbf{C}^n$. For obvious reasons (replacing H by $-H$ if needed) we can assume that H is positive semidefinite. Hence, in what follows, all the results are stated for positive semidefinite H.

Let $H = LL^*$ be any factorization of H. If $\mathcal{X}$ is invariant for H we have

$$(L^*x \mid L^*y) = (LL^*x \mid y) = (Hx \mid y) = 0, \quad x \in \mathcal{X},\ y \in \mathcal{X}^\perp,$$

where $(a \mid b) = b^*a$ for $a, b \in \mathbf{C}^n$. Hence for the subspaces

$$\mathcal{Y}_L = L^*\mathcal{X}, \quad \mathcal{U}_L = L^*\mathcal{X}^\perp$$

we have $\mathcal{Y}_L \subseteq \mathcal{U}_L^\perp$ and $\mathcal{U}_L \subseteq \mathcal{Y}_L^\perp$. This indicates that, at least if L is square and nonsingular, the maximal canonical angle between subspaces $\mathcal{Y}_L$ and $\mathcal{U}_L^\perp$ as well as between $\mathcal{U}_L$ and $\mathcal{Y}_L^\perp$ is zero. Note that $\mathcal{Y}_L$, $\mathcal{Y}_L^\perp$, $\mathcal{U}_L$, and $\mathcal{U}_L^\perp$ are subspaces of $\mathbf{C}^r$ where r is the number of columns of L. Since L can be rectangular and rank deficient, the dimensions of $\mathcal{Y}_L$ and $\mathcal{U}_L^\perp$ ($\mathcal{Y}_L^\perp$ and $\mathcal{U}_L$) can differ. Therefore, we shall use the angle function $\angle(\mathcal{M}_1, \mathcal{M}_2)$ between arbitrary subspaces $\mathcal{M}_1$ and $\mathcal{M}_2$ of $\mathbf{C}^r$, defined by (see [15])

$$\angle(\mathcal{M}_1, \mathcal{M}_2) = \sin^{-1} \min\{\|(I - P_{\mathcal{M}_2})P_{\mathcal{M}_1}\|_2\,,\ \|(I - P_{\mathcal{M}_1})P_{\mathcal{M}_2}\|_2\}.$$

Here $P_{\mathcal{M}_i}$ is the orthogonal projector onto $\mathcal{M}_i$, $i = 1, 2$. From the definition one obtains $\angle(\mathcal{M}_1, \mathcal{M}_2) = \angle(\mathcal{M}_2^\perp, \mathcal{M}_1^\perp)$. Hence for our pairs of subspaces we have $\angle(\mathcal{Y}_L, \mathcal{U}_L^\perp) = \angle(\mathcal{Y}_L^\perp, \mathcal{U}_L) =: \phi_L$. Let us yet show that ϕ_L actually does not depend on L but H. Indeed, let $H = L_1L_1^*$ be another factorization, where L_1 has the same number of columns[1] as L. Then we must have $L_1 = LQ$ for some unitary Q. This implies $\mathcal{Y}_{L_1} = L_1^*\mathcal{X} = Q^*L^*\mathcal{X} = Q^*\mathcal{Y}_L$ and hence $\mathcal{Y}_{L_1}^\perp = Q^*\mathcal{Y}_L^\perp$. In the same way one obtains $\mathcal{U}_{L_1} = Q^*\mathcal{U}_L$ and $\mathcal{U}_{L_1}^\perp = Q^*\mathcal{U}_L^\perp$. Since the angle function is unitarily invariant (see [15]) we have $\angle(\mathcal{Y}_{L_1}, \mathcal{U}_{L_1}^\perp) = \angle(Q^*\mathcal{Y}_L, Q^*\mathcal{U}_L^\perp) = \angle(\mathcal{Y}_L, \mathcal{U}_L^\perp)$ and similarly $\angle(\mathcal{Y}_{L_1}^\perp, \mathcal{U}_{L_1}) = \angle(\mathcal{Y}_L^\perp, \mathcal{U}_L)$. Thus, $\phi_{L_1} = \phi_L =: \phi_H$. Since H is fixed, in what follows we write $\phi = \phi_H$.

Let $X \in \mathbf{C}^{n\times m}$, $X_\perp \in \mathbf{C}^{n\times(n-m)}$ be any orthonormal bases of $\mathcal{X}$, $\mathcal{X}^\perp$. Let $M = X^*HX$, $N = X_\perp^*HX_\perp$ be restrictions of H on $\mathcal{X}$, $\mathcal{X}^\perp$, respectively. Since every transition to new orthonormal bases X', $X'_\perp$ induces unitary matrices U, V such that $X' = XU$, $X'_\perp = X_\perp V$, we have $M' = X'^*HX' = U^*MU$, $N' = X'^*_\perp HX'_\perp = V^*NV$. This shows that the eigenvalues of M and N depend only on H and $\mathcal{X}$.

THEOREM 1.1. *Suppose $H \in \mathbf{C}^{n\times n}$ is Hermitian positive semidefinite and $\mathcal{X}$ is an m-dimensional subspace of $\mathbf{C}^n$. Let $\mu_1 \geq \cdots \geq \mu_m$ and $\eta_{m+1} \geq \cdots \geq \eta_n$ be the eigenvalues of the restrictions of H on $\mathcal{X}$ and $\mathcal{X}^\perp$, respectively. Let $H = LL^*$ and ϕ*

[1] If L_1 has fewer (more) columns than L, an appropriate number of zero columns can be appended to L_1 (L).

$= \angle(L^*\mathcal{X}, (L^*\mathcal{X}^\perp)^\perp)$. *If* $\phi < \pi/2$, *there is an ordering* $\lambda_{j_1}, \ldots, \lambda_{j_n}$ *of the eigenvalues of* H *such that*

$$\begin{aligned}
\frac{|\,\mu_k - \lambda_{j_k}\,|}{|\,\mu_k\,|} &\le \sin\phi, \quad k = 1, \ldots, \dim(L^*\mathcal{X}),\\
\mu_k = \lambda_{j_k} &= 0, \quad k = \dim(L^*\mathcal{X}) + 1, \ldots, m,\\
\frac{|\,\eta_k - \lambda_{j_k}\,|}{|\,\eta_k\,|} &\le \sin\phi, \quad k = m+1, \ldots, m + \dim\left(L^*\mathcal{X}^\perp\right),\\
\eta_k = \lambda_{j_k} &= 0, \quad k = m + \dim\left(L^*\mathcal{X}^\perp\right) + 1, \ldots, n.
\end{aligned}$$

Proof. Let us select orthonormal bases X, $X_\perp$ of $\mathcal{X}$, $\mathcal{X}^\perp$, respectively, such that

$$M = X^*HX = \begin{bmatrix} \mathbf{O} & \\ & \Lambda_1 \end{bmatrix} \begin{matrix} m - r_M \\ r_M \end{matrix}, \quad N = X_\perp^* H X_\perp = \begin{bmatrix} \Lambda_2 & \\ & \mathbf{O} \end{bmatrix} \begin{matrix} r_N \\ n - m - r_N \end{matrix},$$

where $r_M = \operatorname{rank}(M)$, $r_N = \operatorname{rank}(N)$, $\Lambda_1 = \operatorname{diag}(\mu_1, \ldots, \mu_{r_M})$ and $\Lambda_2 = \operatorname{diag}(\eta_{m+1}, \ldots, \eta_{m+r_N})$. Since $[X, X_\perp]^* H [X, X_\perp]$ is positive semidefinite, we have

$$[X, X_\perp]^* H [X, X_\perp] \begin{bmatrix} \mathbf{O} & & \\ & \hat{H} & \\ & & \mathbf{O} \end{bmatrix} \begin{matrix} m - r_M \\ r_M + r_N \\ n - m - r_N \end{matrix}, \quad \hat{H} = \begin{bmatrix} \Lambda_1 & K^* \\ K & \Lambda_2 \end{bmatrix} \begin{matrix} r_M \\ r_N \end{matrix}.$$

Note that $\hat{H}$ has full rank. Therefore, $\hat{H}$ and $\Lambda_1 \oplus \Lambda_2$ are positive definite. By the inertia theorem [12, Theorem 4.1] we conclude that $\hat{H}_S = (\Lambda_1 \oplus \Lambda_2)^{-\frac{1}{2}} \hat{H} (\Lambda_1 \oplus \Lambda_2)^{-\frac{1}{2}}$ is also positive definite. Hence

$$\hat{H} = (\Lambda_1 \oplus \Lambda_2)^{\frac{1}{2}} \hat{H}_S^{\frac{1}{2}} \hat{H}_S^{\frac{1}{2}} (\Lambda_1 \oplus \Lambda_2)^{\frac{1}{2}}$$

and

$$\hat{H}' = \hat{H}_S^{\frac{1}{2}} (\Lambda_1 \oplus \Lambda_2)^{\frac{1}{2}} (\Lambda_1 \oplus \Lambda_2)^{\frac{1}{2}} \hat{H}_S^{\frac{1}{2}} = \hat{H}_S^{\frac{1}{2}} (\Lambda_1 \oplus \Lambda_2) \hat{H}_S^{\frac{1}{2}}$$

have the same eigenvalues, counting multiplicities. Since $\operatorname{rank}(H) = r_M + r_N$, they are exactly the positive eigenvalues of H. Since $\hat{H}'$ is related to $\Lambda_1 \oplus \Lambda_2$ via the congruence transformation with $\hat{H}_S^{\frac{1}{2}}$ we can apply Ostrowski's theorem (see [9] or [7, Theorem 4.5.9]) to $\hat{H}'$. We obtain

$$\lambda_i(\hat{H}') = \theta_i \lambda_i (\Lambda_1 \oplus \Lambda_2), \quad i = 1, \ldots, n \tag{5}$$

with

$$1 - \|K_S\|_2 \le \lambda_{r_M + r_N}(\hat{H}_S) \le \theta_i \le \lambda_1(\hat{H}_S) \le 1 + \|K_S\|_2, \tag{6}$$

where $K_S = \Lambda_2^{-\frac{1}{2}} K \Lambda_1^{-\frac{1}{2}}$. Here $\lambda_i(\cdot)$ denotes the ith largest eigenvalue of a matrix. Note, if we prove

$$\|K_S\|_2 = \sin\phi, \tag{7}$$

then the relations (5) and (6) will imply all the assertions of the theorem.

To prove (7) we use partitions: $X = [X_1, X_2]$, $X_\perp = [X_{\perp,1}, X_{\perp,2}]$, where $X_2 \in \mathbf{C}^{n\times r_M}$, $X_{\perp,1} \in \mathbf{C}^{n\times r_N}$. Then we have

$$K_S = \Lambda_2^{-\frac{1}{2}} K \Lambda_1^{-\frac{1}{2}} = \Lambda_2^{-\frac{1}{2}} X_{\perp,1}^* H X_2 \Lambda_1^{-\frac{1}{2}} = \Lambda_2^{-\frac{1}{2}} X_{\perp,1}^* L L^* X_2 \Lambda_1^{-\frac{1}{2}} \tag{8}$$
$$= \left(L^* X_{\perp,1} \Lambda_2^{-\frac{1}{2}}\right)^* \left(L^* X_2 \Lambda_1^{-\frac{1}{2}}\right) = U^* Y.$$

Note that

$$Y^*Y = \Lambda_1^{-\frac{1}{2}} X_2^* L L^* X_2 \Lambda_1^{-\frac{1}{2}} = \Lambda_1^{-\frac{1}{2}} X_2^* H X_2 \Lambda_1^{-\frac{1}{2}} = \Lambda_1^{-\frac{1}{2}} \Lambda_1 \Lambda_1^{-\frac{1}{2}} = I_{r_M},$$
$$U^*U = \Lambda_2^{-\frac{1}{2}} X_{\perp,1}^* L L^* X_{\perp,1} \Lambda_2^{-\frac{1}{2}} = \Lambda_2^{-\frac{1}{2}} X_{\perp,1}^* H X_{\perp,1} \Lambda_2^{-\frac{1}{2}} = \Lambda_2^{-\frac{1}{2}} \Lambda_2 \Lambda_2^{-\frac{1}{2}} = I_{r_N},$$
$$\mathcal{R}(Y) = \mathcal{R}\left(L^* X_2 \Lambda_1^{-\frac{1}{2}}\right) = \mathcal{R}(L^* X_2) \subseteq \mathcal{R}(L^* X) =: \mathcal{Y}_L,$$
$$\mathcal{R}(U) = \mathcal{R}\left(L^* X_{\perp,1} \Lambda_2^{-\frac{1}{2}}\right) = \mathcal{R}(L^* X_{\perp,1}) \subseteq \mathcal{R}(L^* X_\perp) =: \mathcal{U}_L.$$

Since $0 = X_1^* H X_1 = X_1^* L L^* X_1 = (L^* X_1)^* L^* X_1$ we have $L^* X_1 = 0$. Similarly, we obtain $L^* X_{\perp,2} = 0$. Thus $\dim(\mathcal{R}(L^* X_2)) = \dim(\mathcal{Y}_L)$, $\dim(\mathcal{R}(L^* X_{\perp,1})) = \dim(\mathcal{U}_L)$, whence $\mathcal{R}(Y) = \mathcal{Y}_L$ and $\mathcal{R}(U) = \mathcal{U}_L$. We have shown that Y and U are orthonormal bases for $\mathcal{Y}_L$ and $\mathcal{U}_L$, respectively. Since $\mathcal{Y}_L = L^*\mathcal{X}$, $\mathcal{U}_L = L^*\mathcal{X}^\perp$, we have

$$\sin\phi = \sin\angle\left(\mathcal{Y}_L, \mathcal{U}_L^\perp\right) = \min\left\{\|P_{\mathcal{U}_L} P_{\mathcal{Y}_L}\|_2\ ,\ \|P_{\mathcal{Y}_L^\perp} P_{\mathcal{U}_L^\perp}\|_2\right\}.$$

Using [15, Section 2] we can choose an orthogonal basis of $\mathbf{C}^n$ such that matrix representations of (the linear operators defined by) $P_{\mathcal{U}_L} P_{\mathcal{Y}_L}$, $P_{\mathcal{Y}_L^\perp} P_{\mathcal{U}_L^\perp}$ with respect to that basis are block diagonal matrices. Since $\sin\phi < 1$, the nonzero diagonal blocks are of order two. Furthermore, if $\cos\xi_1 \geq \cdots \geq \cos\xi_\ell$ are the nonzero singular values of K_S, then the nontrivial diagonal blocks of $P_{\mathcal{U}_L} P_{\mathcal{Y}_L}$ and $P_{\mathcal{Y}_L^\perp} P_{\mathcal{U}_L^\perp}$ in the new basis have forms

$$\cos\xi_i \begin{bmatrix} \cos\xi_i & \sin\xi_i \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \cos\xi_i \begin{bmatrix} 0 & -\sin\xi_i \\ 0 & \cos\xi_i \end{bmatrix},$$

respectively. Hence $\sin\phi = \cos\xi_1 = \|K_S\|_2$. This proves the relation (7) and completes the proof of the theorem. □

We call the reader's attention to the beautiful dualities in Theorems 0.4 and 1.1. Theorem 0.4 can be applied to H^{-1} with the same right-hand side in (4). This is in accordance with the fact that H and H^{-1} have the same invariant subspaces. Similarly, in Theorem 1.1 the subspace $\mathcal{X}$ can be replaced by $\mathcal{X}^\perp$ without changing the bound. This corresponds to the fact that $\mathcal{X}$ and $\mathcal{X}^\perp$ are complementary and one is H invariant iff the other is such. In fact, in the case of positive definite H, Theorem 0.4 follows from Theorem 1.1, because in that case $(L^*\mathcal{X}^\perp)^\perp = L^{-1}\mathcal{X}$. In the semidefinite case one can easily show that $(L^*\mathcal{X}^\perp)^\perp = \{x\ :\ Lx \in \mathcal{X}\}$.

2. Quadratic residual bounds. In this section we show how to replace the linear bound of Theorem 1.1 by a bound of order $\sin^2\phi$. Our estimate will differ from that of Theorem 0.2 because we use the relative gap in the spectrum and estimate the relative distance between the spectrum of M and the matching part of the spectrum of H. Note an important restriction in Theorem 0.2. It lies in the definitions of δ_0 and δ: if an eigenvalue of H is approximated by some μ_j, then the whole eigenspace of H corresponding to that eigenvalue has to be approximated by some subspace of

$\mathcal{X}$. Such a condition is unlikely to be met in applications and fortunately it can be removed.

In the following theorem which refines Theorem 1.1, $\sigma_{\min}(\cdot)$ denotes the smallest singular value of a matrix and $\|\cdot\|$ denotes any unitarily invariant matrix norm. We let $\operatorname{Sin}\Phi$ denote the diagonal matrix with sines of the canonical angles between $L^*\mathcal{X}$ and $\left(L^*\mathcal{X}^\perp\right)^\perp$ as diagonal elements. It is actually the diagonal matrix in the singular value decomposition of $K_S = U^*Y$ from the proof of Theorem 1.1. To simplify notation, for a given nonzero eigenvalue λ of H we shall choose the bases X and $X^\perp$ so that

$$\Lambda_1 = \Xi_\lambda \oplus \hat{\Xi}_\lambda\,, \quad \Lambda_2 = \Omega_\lambda \oplus \hat{\Omega}_\lambda, \tag{9}$$

where the diagonals of Ξ_λ and Ω_λ approximate λ in the sense of Theorem 1.1. Note that the diagonals of Λ_1 and Λ_2 need not be in the monotone ordering anymore.

THEOREM 2.1. *Let H, $\mathcal{X}$ and $\phi < \pi/2$ be as in Theorem* 1.1. *Let $\lambda > 0$ be an eigenvalue of H of multiplicity $n(\lambda)$. Let the orthonormal bases of $\mathcal{X}$ and $\mathcal{X}^\perp$ be so chosen that* (9) *holds. Suppose there exist constants $\alpha > \sin\phi$ and $\beta > \sin\phi$ such that*

$$\|\lambda\Xi_\lambda^{-1} - I\|_2 \le \sin\phi, \quad \sigma_{\min}(\lambda\hat{\Xi}_\lambda^{-1} - I) \ge \alpha, \tag{10}$$

$$\|\lambda\Omega_\lambda^{-1} - I\|_2 \le \sin\phi, \quad \sigma_{\min}(\lambda\hat{\Omega}_\lambda^{-1} - I) \ge \beta. \tag{11}$$

If $\Xi_\lambda \oplus \Omega_\lambda$ is of order $n(\lambda)$, then

$$\|I - \lambda\Xi_\lambda^{-1}\| \le \frac{1}{1 - \dfrac{\sin^2\phi}{\alpha\beta}} \frac{\|\operatorname{Sin}\Phi\|_2 \|\operatorname{Sin}\Phi\|}{\beta},$$

$$\|I - \lambda\Omega_\lambda^{-1}\| \le \frac{1}{1 - \dfrac{\sin^2\phi}{\alpha\beta}} \frac{\|\operatorname{Sin}\Phi\|_2 \|\operatorname{Sin}\Phi\|}{\alpha}.$$

Proof. Without loss of generality we can assume

$$H = \begin{bmatrix} \Lambda_1 & K^* \\ K & \Lambda_2 \end{bmatrix},$$

where Λ_1 and Λ_2 are given by (9). Otherwise one can work with $\hat{H}$ from the proof of Theorem 1.1. Since λ is fixed we omit it as matrix subscript. By Sylvester's law of inertia, the matrix

$$H_S(\lambda) = (\Lambda_1 \oplus \Lambda_2)^{-\frac{1}{2}}(H - \lambda I)(\Lambda_1 \oplus \Lambda_2)^{-\frac{1}{2}}$$

has rank $n - n(\lambda)$. It has the following block structure:

$$H_S(\lambda) = \begin{bmatrix} I - \lambda\Xi^{-1} & \mathbf{O} & (K_S^{(1,1)})^* & (K_S^{(2,1)})^* \\ \mathbf{O} & I - \lambda\hat{\Xi}^{-1} & (K_S^{(1,2)})^* & (K_S^{(2,2)})^* \\ K_S^{(1,1)} & K_S^{(1,2)} & I - \lambda\Omega^{-1} & \mathbf{O} \\ K_S^{(2,1)} & K_S^{(2,2)} & \mathbf{O} & I - \lambda\hat{\Omega}^{-1} \end{bmatrix}.$$

For technical reasons we replace $H_S(\lambda)$ by a similar matrix

$$\hat{H}_S(\lambda) = \Pi^\tau H_S(\lambda)\Pi = \begin{bmatrix} I - \lambda\Xi^{-1} & (K_S^{(1,1)})^* & \mathbf{O} & (K_S^{(2,1)})^* \\ K_S^{(1,1)} & I - \lambda\Omega^{-1} & K_S^{(1,2)} & \mathbf{O} \\ \mathbf{O} & (K_S^{(1,2)})^* & I - \lambda\hat{\Xi}^{-1} & (K_S^{(2,2)})^* \\ K_S^{(2,1)} & \mathbf{O} & K_S^{(2,2)} & I - \lambda\hat{\Omega}^{-1} \end{bmatrix},$$

where Π denotes an appropriate permutation matrix. Since the spectral norm of a submatrix is not larger than the norm of the whole matrix, the assumptions (10) and (11) imply

$$\begin{aligned}(12)\qquad \sigma_{\min}((I-\lambda\hat{\Xi}^{-1})\oplus(I-\lambda\hat{\Omega}^{-1})) &\geq \min\{\alpha,\beta\} > \sin\phi \\ &= \|\operatorname{Sin}\Phi\|_2 \geq \|K_S\|_2 \geq \max_{1\leq i,j\leq 2}\|K_S^{(i,j)}\|_2,\end{aligned}$$

where $K_S = \Lambda_2^{-\frac{1}{2}} K \Lambda_1^{-\frac{1}{2}}$. Hence the matrix

$$C = \begin{bmatrix} I-\lambda\hat{\Xi}^{-1} & (K_S^{(2,2)})^* \\ K_S^{(2,2)} & I-\lambda\hat{\Omega}^{-1}\end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22}\end{bmatrix},\ C_{12} = C_{21}^*$$

and its diagonal blocks C_{11} and C_{22} are nonsingular. Therefore (see [7, Section 0.7.3])

$$C^{-1} = \begin{bmatrix} \left[C_{11}-C_{12}C_{22}^{-1}C_{21}\right]^{-1} & C_{11}^{-1}C_{12}\left[C_{21}C_{11}^{-1}C_{12}-C_{22}\right]^{-1} \\ \left[C_{21}C_{11}^{-1}C_{12}-C_{22}\right]^{-1}C_{21}C_{11}^{-1} & \left[C_{22}-C_{21}C_{11}^{-1}C_{12}\right]^{-1}\end{bmatrix},$$

provided that all matrices in brackets are nonsingular. However, this follows since these matrices are (signed) Schur complements of C_{11} and C_{22} in C. By the last assumption, C is of order $n-n(\lambda)$ what is also the rank of $\hat{H}_S(\lambda)$. Since C is nonsingular its Schur complement in $\hat{H}_S(\lambda)$ must be zero (cf. [10, p. 183]). Hence

$$(13)\quad \begin{bmatrix} I-\lambda\Xi^{-1} & (K_S^{(1,1)})^* \\ K_S^{(1,1)} & I-\lambda\Omega^{-1}\end{bmatrix} = \begin{bmatrix} \mathbf{O} & (K_S^{(2,1)})^* \\ K_S^{(1,2)} & \mathbf{O}\end{bmatrix} C^{-1} \begin{bmatrix} \mathbf{O} & (K_S^{(1,2)})^* \\ K_S^{(2,1)} & \mathbf{O}\end{bmatrix}.$$

The rest of the proof is obvious. Indeed, using (13) and the structure of C^{-1} we obtain

$$\begin{aligned} I-\lambda\Xi^{-1} &= (K_S^{(2,1)})^*[I-\lambda\hat{\Omega}^{-1}-K_S^{(2,2)}(I-\lambda\hat{\Xi}^{-1})^{-1}(K_S^{(2,2)})^*]^{-1}K_S^{(2,1)}, \\ I-\lambda\Omega^{-1} &= K_S^{(1,2)}[I-\lambda\hat{\Xi}^{-1}-(K_S^{(2,2)})^*(I-\lambda\hat{\Omega}^{-1})^{-1}K_S^{(2,2)}]^{-1}(K_S^{(1,2)})^*.\end{aligned}$$

Applying an arbitrary unitarily invariant matrix norm to the expressions on the left- and right-hand sides and using its relation to the spectral norm (cf. [12, Theorem 3.9]), we obtain

$$\begin{aligned}\|I-\lambda\Xi^{-1}\| &\leq \frac{\|K_S^{(2,1)}\|_2\|K_S^{(2,1)}\|}{\beta-\dfrac{\|K_S^{(2,2)}\|_2^2}{\alpha}} = \frac{1}{1-\dfrac{\|K_S^{(2,2)}\|_2^2}{\alpha\beta}}\,\frac{\|K_S^{(2,1)}\|_2\|K_S^{(2,1)}\|}{\beta}, \\ \|I-\lambda\Omega^{-1}\| &\leq \frac{\|K_S^{(1,2)}\|_2\|K_S^{(1,2)}\|}{\alpha-\dfrac{\|K_S^{(2,2)}\|_2^2}{\beta}} = \frac{1}{1-\dfrac{\|K_S^{(2,2)}\|_2^2}{\alpha\beta}}\,\frac{\|K_S^{(1,2)}\|_2\|K_S^{(1,2)}\|}{\alpha}.\end{aligned}$$

Since for any unitarily invariant norm (cf. [12, Corollary 3.8])

$$\max_{1\leq i,j\leq 2}\|K_S^{(i,j)}\| \leq \|K_S\| = \|\operatorname{Sin}\Phi\|,$$

the proof is completed. □

If the matrices $\hat{\Xi}_\lambda$ and Ω_λ in Theorem 2.1 are void, i.e., in the proof of Theorem 2.1, $\Lambda_1 = \Xi_\lambda$ and $\Lambda_2 = \hat{\Omega}_\lambda$, then the assertion of the theorem reduces to

$$\|I - \lambda \Xi_\lambda^{-1}\| \leq \frac{\|\sin\Phi\|_2 \|\sin\Phi\|}{\beta}. \tag{14}$$

This is the first assertion of Theorem 2.1 with a simplified bound.

Specifying $\|\cdot\|$ to be the spectral and the Frobenius norm, we obtain the following useful relative a posteriori estimates.

COROLLARY 2.2. *Let $\mathcal{L}$ denote the set of all $\lambda \in \sigma(H)$ for which the assumptions of Theorem 2.1 hold. For $\lambda \in \mathcal{L}$ let Ξ_λ, Ω_λ, α_λ, and β_λ be as in Theorem 2.1. If $\alpha = \min_{\lambda\in\mathcal{L}} \alpha_\lambda$, $\beta = \min_{\lambda\in\mathcal{L}} \beta_\lambda$ then*

$$\max_{\lambda\in\mathcal{L}} \max_{\mu\in\sigma(\Xi_\lambda)} \frac{|\mu-\lambda|}{|\mu|} \leq \frac{1}{1-\dfrac{\sin^2\phi}{\alpha\beta}} \frac{\sin^2\phi}{\beta},$$

$$\max_{\lambda\in\mathcal{L}} \max_{\eta\in\sigma(\Omega_\lambda)} \frac{|\eta-\lambda|}{|\eta|} \leq \frac{1}{1-\dfrac{\sin^2\phi}{\alpha\beta}} \frac{\sin^2\phi}{\alpha},$$

$$\sqrt{\sum_{\lambda\in\mathcal{L}} \sum_{\mu\in\sigma(\Xi_\lambda)} \left(\frac{\mu-\lambda}{\mu}\right)^2} \leq \frac{1}{1-\dfrac{\sin^2\phi}{\alpha\beta}} \frac{\|\sin\Phi\|_F^2}{\beta},$$

$$\sqrt{\sum_{\lambda\in\mathcal{L}} \sum_{\eta\in\sigma(\Omega_\lambda)} \left(\frac{\eta-\lambda}{\eta}\right)^2} \leq \frac{1}{1-\dfrac{\sin^2\phi}{\alpha\beta}} \frac{\|\sin\Phi\|_F^2}{\alpha}.$$

Remark 2.3. One can easily check that the assumptions (10) and (11) of Theorem 2.1 are satisfied if, e.g., for some $\lambda \in \sigma(H)$

$$\sin\phi \leq \frac{1}{3} \min_{\lambda_i\neq\lambda} \frac{|\lambda_i-\lambda|}{\lambda_i+\lambda} \equiv \delta_\lambda.$$

In such a case α and β in Theorem 2.1 can be replaced by δ_λ.

Example 2.4. Let

$$H = \begin{bmatrix} 10^{10} & 1 & 10^{-13} \\ 1 & 2\cdot 10^{-5} & 10^{-7} \\ 10^{-13} & 10^{-7} & 10^{-5} \end{bmatrix}, \quad X = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad X^\perp = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then

$$M = \begin{bmatrix} 2\cdot 10^{-5} \end{bmatrix}, \quad R = \begin{bmatrix} 10^{-7} & 0 & 1 \end{bmatrix}^*.$$

Thus, the result of Theorem 0.1 is not useful because $\|R\|_2 \approx 1$. Since

$$H_S = \mathrm{diag}\,(H_{ii}^{-\frac{1}{2}}) H \mathrm{diag}\,(H_{ii}^{-\frac{1}{2}}) = I + E, \quad \|E\|_\infty < 10^{-2},$$

we know from the theory of Barlow and Demmel [1] that

$$1 - 10^{-2} < \frac{H_{jj}}{\lambda_j} < 1 + 10^{-2}, \quad j = 1, 2, 3,$$

where $\lambda_1 \geq \lambda_2 \geq \lambda_3$ are the eigenvalues of H. This means that in this example the separation δ from Theorem 0.2 is of order 10^{-5} and that Theorem 0.2 is not applicable. On the other hand, both Theorem 0.4 and Theorem 1.1 ensure that for some $j_0 \in \{1, 2, 3\}$ (check that $j_0 = 2$) holds $|\lambda_{j_0} - 2 \cdot 10^{-5}|/\lambda_{j_0} < 7.5 \cdot 10^{-3}$. Note that $N = (X^{\perp})^* H X^{\perp}$ is scaled diagonally dominant in the sense of [1] and its diagonal elements approximate the eigenvalues of N to 15 significant digits. Hence one can check that the eigenvalue λ_{j_0} is well separated (in the relative sense) from the spectrum of N. An easy calculation shows that we can take $\beta = 0.9$. Since ϕ is the angle between two one-dimensional subspaces, the relation (14) and an easy calculation yield $|\lambda_{j_0} - 2 \cdot 10^{-5}|/\lambda_{j_0} < 6.2 \cdot 10^{-5}$. Note that $|\lambda_{j_0} - 2 \cdot 10^{-5}|/\lambda_{j_0} \approx |\lambda_{j_0} - 2 \cdot 10^{-5}|/2 \cdot 10^{-5} \approx 4.5 \cdot 10^{-5}$.

Acknowledgment. We would like to thank Dr. N. J. Higham, Manchester, and anonymous referees for detailed and constructive comments. We also thank Professor Ji-guang Sun, Umeå, for calling our attention to [13].

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] Z.-H. Cao, J.-J. Xie, and R.-C. Li, *A sharp version of Kahan's theorem on clustered eigenvalues*, Computer Science Division Technical Report UCB//CSD-94-857, University of California, Berkeley, CA 94720, 1994.
[3] F. Chatelin, *Spectral Approximation of Linear Operators*, Academic Press, New York, 1983.
[4] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.
[5] Z. Drmač, *Computing the Singular and the Generalized Singular Values*, Ph.D. thesis, Lehrgebiet Mathematische Physik, Fernuniversität Hagen, Germany, 1994.
[6] Z. Drmač, *On relative residual bounds for the eigenvalues of a Hermitian matrix*, Linear Algebra Appl., 244 (1996), pp. 155–163.
[7] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, London, 1990.
[8] W. Kahan, *Inclusion theorems for clusters of eigenvalues of Hermitian matrices*, Technical report, Computer Science Department, University of Toronto, Toronto, Ontario, Canada, 1967.
[9] A. M. Ostrowski, *A quantitative formulation of Sylvester's law of inertia*, Proc. Nat. Acad. Sci. U.S.A., 45 (1959), pp. 740–744.
[10] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[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] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
[13] J.-G. Sun, *Eigenvalues of Rayleigh quotient matrices*, Numer. Math., 59 (1991), pp. 603–614.
[14] K. Veselić and I. Slapničar, *Floating-point perturbations of Hermitian matrices*, Linear Algebra Appl., 195 (1993), pp. 81–116.
[15] P. A. Wedin, *On angles between subspaces of a finite dimensional inner product space*, in Matrix Pencils Proceedings of a Conference, Springer-Verlag, Berlin, 1982.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 30–36, January 1997

003

GMRES VS. IDEAL GMRES*

KIM-CHUAN TOH†

Abstract. The GMRES algorithm minimizes $\|p(A)b\|$ over polynomials p of degree n normalized at $z = 0$. The ideal GMRES problem is obtained if one considers minimization of $\|p(A)\|$ instead. The ideal problem forms an upper bound for the worst-case true problem, where the GMRES norm $\|p_b(A)b\|$ is maximized over b. In work not yet published, Faber, Joubert, Knill, and Manteuffel have shown that this upper bound need not be attained, constructing a 4×4 example in which the ratio of the true to ideal GMRES norms is 0.9999. Here, we present a simpler 4×4 example in which the ratio approaches zero when a certain parameter tends to zero. The same example also leads to the same conclusion for Arnoldi vs. ideal Arnoldi norms.

Key words. GMRES, ideal GMRES, Arnoldi, ideal Arnoldi

AMS subject classifications. 65F10, 49K35

PII. S089547989427909X

1. Introduction.

The GMRES algorithm [7] is an iterative method for solving non-hermitian linear systems $Ax = b$ ($A \in \mathbb{C}^{N\times N}$, $b \in \mathbb{C}^N$). Throughout this paper, $\mathbb{C}^N$ is given the 2-norm $\|\cdot\|$ and $\mathbb{C}^{N\times N}$ is given the corresponding induced matrix norm. Each step (say the nth) of the GMRES algorithm is mathematically equivalent[1] to minimizing $\|p(A)b\|$ over the polynomials in P_n, where

$$P_n = \{\text{polynomials of degree} \leq n \text{ with } p(0) = 1\}.$$

For each b, the GMRES polynomial (denoted by p_b) exists and is unique if $\|p_b(A)b\| > 0$.

How fast a GMRES iteration converges, i.e., how fast $\|p_b(A)b\|$ converges to zero as n increases, depends on the matrix A and the vector b. In practice, however, unless b has special properties, it appears to be usually A that predominantly determines the convergence rate. To understand how the GMRES convergence rate depends on A without the complicating effect of the right-hand side vector, Greenbaum and Trefethen [5] introduced the "ideal GMRES matrix approximation problem": minimization of $\|p(A)\|$ over polynomials in the same class P_n. The "ideal GMRES polynomial," which we will denote by p_*, exists and is unique so long as $\|p_*(A)\| > 0$. To avoid possible confusion, we will refer to GMRES as true GMRES.

The ideal GMRES convergence curve forms an upper bound for the true GMRES convergence curves in the sense that for each n,

$$\max_{b\in\mathbb{C}^N,\ \|b\|=1} \|p_b(A)b\| \leq \|p_*(A)\|. \tag{1.1}$$

This inequality is actually an equality for many matrices, including normal matrices [3], [4], triangular Toeplitz matrices with $p_*(z) = 1$ [2], and matrices A whose ideal GMRES matrix $p_*(A)$ has a simple maximal singular value [5]. It is also an equality for

*Received by the editors December 23, 1994; accepted for publication (in revised form) by R. Freund December 18, 1995.

http://www.siam.org/journals/simax/18-1/27909.html

†Department of Mathematics, National University of Singapore, 10 Kent Ridge Crescent, Singapore 0511 (mattohkc@leonis.nus.sg). This author was supported by a National University of Singapore Graduate Scholarship, NSF grant DMS-9116110, and DOE grant DE-FG02-94ER25199.

[1]We have assumed, without loss of generality, that the initial guess for the iteration is $x_0 = 0$.

arbitrary matrices at step $n = 1$ [3],[4]. Positive results such as these led Greenbaum and Trefethen [5] to conjecture that (1.1) was an equality, i.e., "the ideal GMRES bound is attained," for every matrix A. However, at the 1994 Colorado Conference on Iterative Methods at Breckenridge, Colorado, Faber, Joubert, Knill, and Manteuffel presented a counterexample to this conjecture [2]. Their example is a dense 4×4 matrix constructed using the theory of generalized fields of values, where the inequality (1.1) is strict at step $n = 3$. The degree-3 ideal GMRES polynomial for their example is $p_*(z) = 1$; hence $\|p_*(A)\| = 1$. The corresponding quantity on the left-hand side of (1.1) is 0.99988.

We have found a simpler (bidiagonal) family of 4×4 matrices that can achieve arbitrarily small ratio when a certain parameter in the family tends to zero. The purpose of this short paper is to present this example and speculate briefly on its significance.

2. The counterexample: Mathematical proof. Our counterexample is the 4×4 matrix

$$A = \begin{pmatrix} 1 & \epsilon & & \\ & -1 & c/\epsilon & \\ & & 1 & \epsilon \\ & & & -1 \end{pmatrix}, \quad \epsilon > 0, \quad 0 < c < 2. \tag{2.1}$$

We would like to note that the parameter c in the example is not crucial to establishing our goal, namely, to show that the worst-case true and ideal GMRES norms in (1.1) differ. However, it gives us the freedom to construct examples with an ideal GMRES norm anywhere between zero and one. For simplicity, the reader can assume c to be one.

THEOREM 2.1. *For the matrix A of* (2.1)*, the degree-*3 *ideal GMRES polynomial is*

$$p_*(z) = 1 + (\alpha - 1)z^2 \tag{2.2}$$

with

$$\alpha = \frac{2c^2}{4 + c^2}. \tag{2.3}$$

The corresponding matrix is

$$p_*(A) = \begin{pmatrix} \alpha & 0 & \gamma & \\ & \alpha & 0 & \gamma \\ & & \alpha & 0 \\ & & & \alpha \end{pmatrix},$$

where

$$\gamma = (\alpha - 1)c, \tag{2.4}$$

with norm

$$\|p_*(A)\| = \frac{4c}{4 + c^2}. \tag{2.5}$$

Proof. Since A is real, we have $\|p(A)\| = \|\bar{p}(A)\|$ for any p. Uniqueness of the ideal GMRES polynomial then implies that $p_*(z) = \bar{p}_*(z)$, i.e., the coefficients of $p_*(z)$ are real. Next we observe that A^T is unitarily similar to $-A$ via the matrix

$$Q = \begin{pmatrix} & & & -1 \\ & & 1 & \\ & -1 & & \\ 1 & & & \end{pmatrix};$$

i.e., $-A = QA^TQ^{-1}$. This implies that $\|p(-A)\| = \|p(A^T)\| = \|p(A)\|$ for any p. By uniqueness, again, we have $p_*(-z) = p_*(z)$; i.e., p_* is even.

Now consider polynomials of the form (2.2), viewing α as a real parameter. For a given pair of ϵ and c, we can find the singular values of $p(A)$ analytically as a function of α:

$$\sigma^2_{\max}(\alpha) = \frac{1}{2}\left(2\alpha^2 + \gamma^2 + |\gamma|\sqrt{4\alpha^2+\gamma^2}\right), \tag{2.6}$$

$$\sigma^2_{\min}(\alpha) = \frac{1}{2}\left(2\alpha^2 + \gamma^2 - |\gamma|\sqrt{4\alpha^2+\gamma^2}\right), \tag{2.7}$$

with γ related to α and c by (2.4). Each of these singular values has multiplicity two. The value of α corresponding to the ideal GMRES polynomial $p_*(z)$ is the value for which $\sigma^2_{\max}(\alpha)$ is minimum. Now we have a calculus problem; we can simply differentiate (2.6) with respect to α and set the derivative to zero. This gives us the formula (2.3) for α as a function of c; we omit the details. The corresponding singular values of $p_*(A)$ are

$$\sigma_{\max} = \frac{4c}{4+c^2}, \qquad \sigma_{\min} = \frac{c^3}{4+c^2}. \tag{2.8}$$

A biorthogonal set of basis vectors for the maximal left and right singular spaces of $p_*(A)$ is

$$U_1 = \begin{pmatrix} 0 & 2 \\ 2 & 0 \\ 0 & -c \\ -c & 0 \end{pmatrix}, \qquad V_1 = \begin{pmatrix} 0 & c \\ c & 0 \\ 0 & -2 \\ -2 & 0 \end{pmatrix}. \tag{2.9}$$

In what follows, we will denote the columns of U_1 and V_1, respectively, by u_i $(i = 1, 2)$ and v_i $(i = 1, 2)$. ∎

Remark. The results of Theorem 2.1 are valid only for $0 < c < 2$. However we may extend these results to the case $c = 2$, since $p_*(A)$ is a continuous function of c. For example, by letting c tend to two in (2.2), we have $p_*(z) = 1$ and hence $\|p_*(A)\| = 1$.

It is easily shown that for our matrix A, the worst-case true and ideal GMRES norms differ. Before attempting to quantify this difference, we will show that it exists.

THEOREM 2.2. *Suppose A is given by* (2.1). *Then for any vector $b \in \mathbb{C}^4$, the corresponding degree-3 true GMRES polynomial p_b for A satisfies $\|p_b(A)b\| < \|p_*(A)\|\|b\|$.*

Proof. We prove this by contradiction. Suppose the envelope is attained, i.e., equality holds in (1.1) for some b. It is easily shown that b must be a maximal right singular vector of $p_*(A)$. That is, it lies in the span of v_1 and v_2, and the corresponding

true GMRES polynomial must be $p_*(z)$ itself. Without loss of generality, we can assume that b has the form $b = \beta_1 v_1 + \beta_2 v_2$, where β_1, β_2 are not both zero. Since p_b is a true GMRES polynomial for b, it is readily shown that $p_*(A)b$ must satisfy the orthogonality conditions

$$\langle A^k b,\, p_*(A)b\rangle = 0, \quad k = 1, 2, 3.$$

Noting that $p_*(A)v_i = \sigma_{\max}\, u_i,\ i = 1, 2$ and evaluating the inner products for $k = 1$ and 3 gives

$$-4c\,|\beta_1|^2 + 4c\,|\beta_2|^2 - 4\frac{c}{\epsilon}\,\bar{\beta}_1\beta_2 + 4c\epsilon\,\beta_1\bar{\beta}_2 = 0, \tag{2.10}$$

$$-4\frac{c}{\epsilon}\,\bar{\beta}_1\beta_2 = 0. \tag{2.11}$$

Equation (2.11) implies that $\beta_1 = 0$ or $\beta_2 = 0$. In either case, substitution into (2.10) gives $c = 0$. Since $c \neq 0$, we have a contradiction. ∎

Our larger goal is to show that the worst-case true and ideal GMRES norms do not merely differ but can have a ratio arbitrarily small. For this we can use the following more quantitative argument.

THEOREM 2.3. *Suppose A is given by* (2.1) *with $0 < \epsilon \leq 1$. Then*

$$\max_{\|b\|=1} \|p_b(A)b\| \leq 2(1+c)\sqrt{\epsilon} \;+\; (2+3c)\epsilon. \tag{2.12}$$

Thus for each $0 < c < 2$,

$$\frac{\max_{\|b\|=1} \|p_b(A)b\|}{\|p_*(A)\|} \longrightarrow 0 \quad \text{as} \;\; \epsilon \to 0. \tag{2.13}$$

Proof. We will show that for each $b \in \mathbb{C}^4$ with $\|b\| = 1$, there exists a polynomial $p \in P_n$ such that $\|p(A)b\|$ is less than or equal to the right-hand side of (2.12). Then (2.12) follows from the optimality property of GMRES.

Let $b = (b_1, b_2, b_3, b_4)^T$. We have

$$Ab = \begin{pmatrix} b_1 \\ -b_2 + cb_3/\epsilon \\ b_3 \\ -b_4 \end{pmatrix} + \epsilon \begin{pmatrix} b_2 \\ 0 \\ b_4 \\ 0 \end{pmatrix}, \quad A^2 b = b + c \begin{pmatrix} b_3 \\ b_4 \\ 0 \\ 0 \end{pmatrix},$$

$$A^3 b = Ab + c \begin{pmatrix} b_3 \\ -b_4 \\ 0 \\ 0 \end{pmatrix} + c\epsilon \begin{pmatrix} b_4 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Consider polynomials $p \in P_n$ of the form

$$p(z) = 1 + \xi z - z^2 + (\eta - \xi) z^3.$$

Then

$$p(A)b = \begin{pmatrix} -cb_3 \;+\; (\eta-\xi)cb_3 \;+\; \eta b_1 \\ -cb_4 \;-\; (\eta-\xi)cb_4 \;-\; \eta b_2 \;+\; \eta cb_3/\epsilon \\ \eta b_3 \\ -\eta b_4 \end{pmatrix} \;+\; \epsilon\, r,$$

where

$$r = \eta \begin{pmatrix} b_2 \\ 0 \\ b_4 \\ 0 \end{pmatrix} + (\eta - \xi)c \begin{pmatrix} b_4 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Now we have two cases to consider. For each, we will show that $\|p(A)b\|$ is less than or equal to the right-hand side of (2.12) with appropriately chosen ξ and η.

Case 1. Suppose $|b_3| \geq \sqrt{\epsilon}$. Take $\xi = -1$ and $\eta = 2\epsilon b_4/b_3$. Then $|\eta| \leq 2\sqrt{\epsilon}$ and

$$p(A)b = \eta \begin{pmatrix} b_1 + cb_3 \\ -b_2 - cb_4 \\ b_3 \\ -b_4 \end{pmatrix} + \epsilon\, r.$$

Hence

$$\|p(A)b\| \;\leq\; (1+c)\,|\eta| \;+\; (2+3c)\,\epsilon \;\leq\; 2(1+c)\sqrt{\epsilon} \;+\; (2+3c)\,\epsilon.$$

Case 2. Suppose $|b_3| \leq \sqrt{\epsilon}$. Take $\xi = 1$ and $\eta = \epsilon$. Then

$$p(A)b = c \begin{pmatrix} -2b_3 \\ b_3 \\ 0 \\ 0 \end{pmatrix} + \eta \begin{pmatrix} b_1 + cb_1 \\ -b_2 - cb_4 \\ b_3 \\ -b_4 \end{pmatrix} + \epsilon\, r.$$

Thus

$$\begin{aligned} \|p(A)b\| &\leq \sqrt{5}c\,|b_3| \;+\; (1+c)\,|\eta| + (1+2c)\,\epsilon \\ &\leq 2(1+c)\sqrt{\epsilon} \;+\; (2+3c)\,\epsilon. \qquad \square \end{aligned}$$

Remark. Note that (2.12) in fact holds for all $c > 0$. Since $\|p_*(A)\| = 1$ for all $\epsilon > 0$ when $c = 2$, as a result (2.13) also holds for $c = 2$.

3. The counterexample: Numerical evidence. Theorem 2.3 shows that the ratio of the true to ideal GMRES norms for our matrix A is no greater than order $\sqrt{\epsilon}$ as $\epsilon \to 0$. In fact, numerical experiments indicate that this square root dependence is sharp. We have used the MATLAB optimization routine `fminu` [6] to maximize $\|p_b(A)b\|$ over $b \in \mathbb{C}^4$ with $\|b\| = 1$. To ensure that we have the global maximum for the worst-case true GMRES, numerous trails with different initial guesses are carried out with `fminu`. The ideal GMRES polynomial is computed from (2.2).

Figure 3.1 plots the ratio between the worst-case true GMRES and the ideal GMRES norms for the matrix A of (2.1) with $0 < \epsilon \leq 10$ and $c = 1$. The dashed curve shows an upper bound on the ratio obtained by dividing the right-hand quantity of (2.12) by the ideal GMRES norm of A in (2.5). The slope of the curves in the figure is 0.5.

By extending the matrix A of (2.1) to higher dimensions, say to an even integer N (with ± 1 alternating along the diagonal and ϵ, c/ϵ alternating along the first superdiagonal), we obtain examples where the ideal GMRES envelope is not attained at step $n = N - 1$. For such matrices, again, the ideal GMRES polynomials do not depend on ϵ. We have used codes provided by Michael Overton to compute the ideal GMRES polynomials. Numerical experiments also indicate that the worst-case true GMRES norms at step $n = N - 1$ are no greater than order $\sqrt{\epsilon}$ as $\epsilon \to 0$. Thus the ratio between the worst-case true GMRES and ideal GMRES norms at step $n = N-1$ approaches zero as ϵ tends to zero.

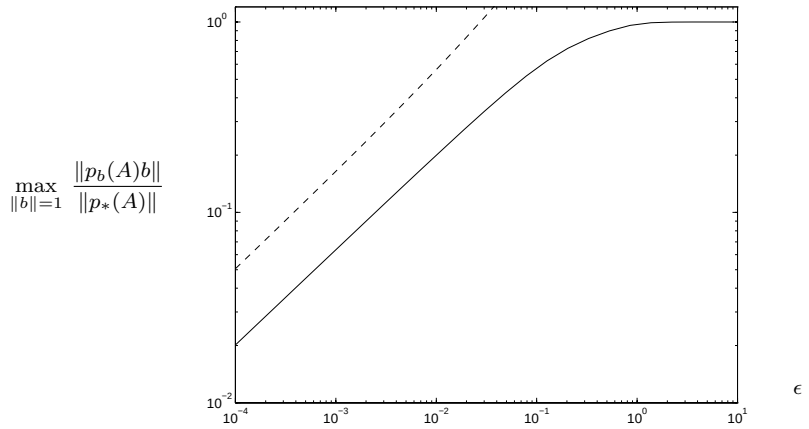


FIG. 3.1. *Ratio between the worst-case true GMRES and ideal GMRES norms at step $n = 3$ for the matrix A of (2) with $c = 1$, as a function of ϵ (numerically computed). The plateau portion of the solid curve is strictly below 1 for all ϵ, by Theorem 2.2. The dashed curve represents the upper bound of Theorem 2.3.*

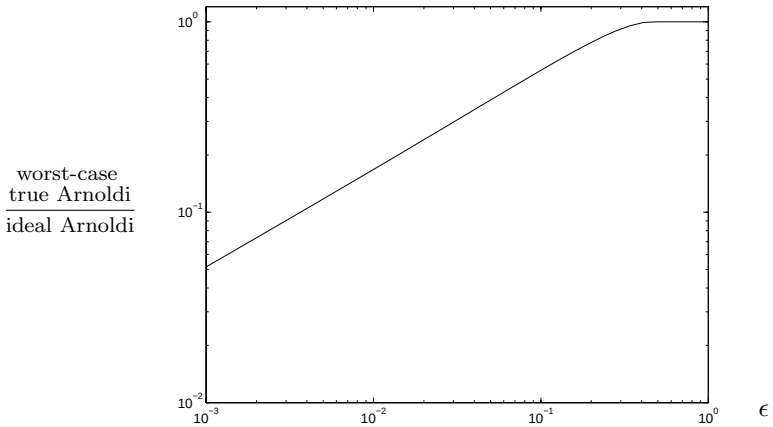


FIG. 4.1. *Analogous plot for true vs. ideal Arnoldi approximation of the same matrix A with $c = 1$. The ratio is exactly 1 for ϵ approximately greater than 0.459.*

4. Discussion. The true and ideal Arnoldi problems are the analogs of the true and ideal GMRES problems, except that the minimizations are over the class of monic polynomials of degree $\leq n$ instead of P_n. Numerical evidence again suggests that for the matrix A of (2.1), the ratio between the worst-case true Arnoldi and the ideal Arnoldi norms at step $n = 3$ approaches zero as ϵ tends to zero. Figure 4.1 plots the ratio associated with the Arnoldi problems for our matrix A with $10^{-3} \leq \epsilon \leq 1$ and $c = 1$.

Finally, we must raise the question of the practical significance of our results. Greenbaum and Trefethen [5], as well as others, have assumed that for most non-symmetric matrix iterations in most applications, convergence rates can be analyzed in terms of a matrix approximation problem. Our result introduces the possibility

that this might not be true. There may be applications in which Krylov subspace iterations perform much better than analysis of matrix approximation problems can explain, and conceivably, such applications might be common. Our guess is that this will not prove to be the case, but it must be admitted that at the moment there is very little evidence one way or the other.

Acknowledgments. The author thanks Anne Greenbaum and Nick Trefethen for many stimulating discussions. Nick Trefethen also carefully read drafts of this paper and suggested numerous improvements. The author also thanks Michael Overton for providing him with MATLAB codes designed to find the minimum largest eigenvalue of functions of symmetric matrices [1]. These codes were used to compute the ideal GMRES and ideal Arnoldi polynomials. Finally, the author thanks one of the referees for pointing out a mistake in the original manuscript submitted.

REFERENCES

[1] F. ALIZADEH, J.-P. A. HAEBERLY, AND M. L. OVERTON, *Primal-dual interior-point methods for semidefinite programming: Convergence rates, stability and numerical results*, Report 721, Computer Science Department, New York University, New York, 1996.

[2] V. FABER, W. JOUBERT, E. KNILL, AND T. MANTEUFFEL, *Minimal residual method stronger than polynomial preconditioning*, in Proc. Colorado Conference on Iterative Methods, Breckenridge, CO, 1994.

[3] W. A. JOUBERT, *A robust GMRES-based adaptive polynomial preconditioning algorithm for nonsymmetric linear systems*, SIAM J. Sci. Comput., 15 (1994), pp. 427–439.

[4] A. GREENBAUM AND L. GURVITS, *Max-min properties of matrix factor norms*, SIAM J. Sci. Comput., 15 (1994), pp. 348–358.

[5] A. GREENBAUM AND L. N. TREFETHEN, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems*, SIAM J. Sci. Comput., 15 (1994), pp. 359–368.

[6] THE MATHWORKS, INC., *Optimization Toolbox*, The MathWorks, Inc., Natick, MA, 1992.

[7] Y. SAAD AND M. H. SCHULTZ, *GMRES: A generalized minimum residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 37–51, January 1997

004

GMRES ON (NEARLY) SINGULAR SYSTEMS*

PETER N. BROWN† AND HOMER F. WALKER‡

Abstract. We consider the behavior of the GMRES method for solving a linear system $Ax = b$ when A is singular or nearly so, i.e., ill conditioned. The (near) singularity of A may or may not affect the performance of GMRES, depending on the nature of the system and the initial approximate solution. For singular A, we give conditions under which the GMRES iterates converge safely to a least-squares solution or to the pseudoinverse solution. These results also apply to any residual minimizing Krylov subspace method that is mathematically equivalent to GMRES. A practical procedure is outlined for efficiently and reliably detecting singularity or ill conditioning when it becomes a threat to the performance of GMRES.

Key words. GMRES method, residual minimizing methods, Krylov subspace methods, iterative linear algebra methods, singular or ill-conditioned linear systems

AMS subject classification. 65F10

PII. S0895479894262339

1. Introduction.

The generalized minimal residual (GMRES) method of Saad and Schultz [16] is widely used for solving a general linear system

$$Ax = b, \qquad A \in \mathbb{R}^{n\times n}, \tag{1.1}$$

and its behavior is well understood when A is nonsingular. Our purpose here is to examine the behavior of GMRES when A is singular or nearly so, i.e., ill conditioned, and to formulate practically effective ways of recognizing singularity or ill conditioning when it might significantly affect the performance of the method.

Abstractly, GMRES begins with an initial approximate solution x_0 and initial residual $r_0 \equiv b - Ax_0$ and characterizes the kth approximate solution as $x_k = x_0 + z_k$, where z_k solves

$$\min_{z\in\mathcal{K}_k} \|b - A(x_0 + z)\|_2 = \min_{z\in\mathcal{K}_k} \|r_0 - Az\|_2. \tag{1.2}$$

Here, $\mathcal{K}_k$ is the kth Krylov subspace determined by A and r_0, defined by

$$\mathcal{K}_k \equiv \text{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}.$$

There are a number of ways of implementing GMRES, but in each one generates a basis of $\mathcal{K}_k$ and then replaces (1.2) by an unconstrained k-dimensional least-squares

* Received by the editors January 31, 1994; accepted for publication (in revised form) by R. Freund February 17, 1996.

http://www.siam.org/journals/simax/18-1/26233.html

† Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA 94550 (pnbrown@llnl.gov). The research of this author was supported in part by the Applied Mathematical Sciences subprogram of the Office of Scientific Computing, U.S. Dept. of Energy, by Lawrence Livermore National Laboratory under contract W-7405-ENG-48.

‡ Department of Mathematics and Statistics, Utah State University, Logan, UT 84322-3900 (walker@math.usu.edu). The work of this author was supported in part by U.S. Air Force Office of Scientific Research grant AFOSR-91-0294, U.S. Dept. of Energy grants DE-FG02-92ER25136 and DE-FG03-94ER25221, and National Science Foundation grant DMS-9400217, all with Utah State University. It was done in part during visits to the Computing and Mathematics Research Division, Lawrence Livermore National Laboratory, and the Center for Research on Parallel Computation, Rice University.

problem. We shall not be more specific about the basis generating process at this point, except to assume that it successfully generates a basis if and only if $\dim \mathcal{K}_k = k$, where "dim" denotes dimension.

Note that, trivially, $\dim A(\mathcal{K}_k) \leq \dim \mathcal{K}_k \leq k$ for each k. We shall say that GMRES *does not break down* at the kth step if $\dim A(\mathcal{K}_k) = k$. In this case, $\dim A(\mathcal{K}_k) = \dim \mathcal{K}_k$ and, hence, (1.2) has a unique solution. Furthermore, since $\dim \mathcal{K}_k = k$, a basis of $\mathcal{K}_k$ is successfully generated and the k-dimensional least-squares problem also has a unique solution. This definition addresses two distinct kinds of breakdown: *rank deficiency of the least-squares problem* (1.2), which occurs when $\dim A(\mathcal{K}_k) < \dim \mathcal{K}_k$, and *degeneracy of* $\mathcal{K}_k$, which occurs when $\dim \mathcal{K}_k < k$. The definition is intended to focus on essential breakdown of the method, as opposed to breakdown associated with any particular implementation or ancillary algorithm used in it. Note that if $\dim A(\mathcal{K}_k) < k$ for some k, then $\mathcal{K}_j = \mathcal{K}_k$ for all $j \geq k$ and no further improvement is possible, even if subsequent $z_j \in \mathcal{K}_j$ are well defined in some way.

For perspective, we recall that Proposition 2, p. 865, of [16] ensures that, if A is nonsingular, then GMRES does not break down until the solution of (1.1) has been found. Breakdown in [16, Prop. 2, p. 865] is associated specifically with breakdown of the Arnoldi process used in the GMRES implementation in [16], but the statement remains true with our definition (see section 2 below).

In contrast to the nonsingular case, anything may happen when A is singular. Example 1.1 below shows that GMRES may break down before getting anywhere at all, even when the system has a solution, or it may determine a least-squares solution[1] or the pseudoinverse solution[2] without breaking down. Example 1.2 shows that even if a least-squares solution or the pseudoinverse solution is reached, this may not be evident from the behavior of GMRES; indeed, GMRES may continue for a number of additional steps without breakdown (or further progress).

Example 1.1. Suppose that

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad x_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Then $r_0 = (1,0)^T$ and $Ar_0 = (0,0)^T$, and GMRES breaks down at the first step. Note that x_0 is not a (least-squares) solution. If A is changed to

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix},$$

then, for the same b and x_0, we have $r_0 = (1,0)^T = Ar_0$, and GMRES determines without breakdown $x_1 = (1,0)^T$, which is a least-squares solution but not the pseudoinverse solution. If we also change b to $b = (1,1)^T$, then, for the same x_0, we have $r_0 = (1,1)^T$ and $Ar_0 = (2,0)^T$, and GMRES determines without breakdown $x_1 = (1/2, 1/2)^T$, which is the pseudoinverse solution. Note that $\dim A(\mathcal{K}_2) = 1$ in these last two cases, so GMRES breaks down at the step after the least-squares or pseudoinverse solution has been found.

Example 1.2. For arbitrary n, let A be the "shift" operator with ones on the first subdiagonal and zeros elsewhere. Then for $b = (1,0,\ldots,0)^T$ and $x_0 = (0,\ldots,0)^T$,

[1] An $x \in \mathbb{R}^n$ for which $\|b - Ax\|_2$ is minimal.
[2] The least-squares solution x such that $\|x\|_2$ is minimal.

x_0 itself is the pseudoinverse solution, but GMRES proceeds without breakdown (or progress) until the nth step, at which point it breaks down with $\dim A(\mathcal{K}_n) = n - 1$.

In section 2 below, we explore the theoretical behavior of GMRES when A is singular and, in particular, determine circumstances in which the GMRES iterates converge without breakdown to a least-squares solution or the pseudoinverse solution of (1.1). We also discuss the conditioning of the least-squares problem (1.2) prior to breakdown, since this is crucial to the practical performance of the method. The results in section 2 apply not only to GMRES but also to any mathematically equivalent method, i.e., any method that takes steps characterized by the residual minimizing property (1.2). (See [8, sect. 2.4] for a discussion of mathematically equivalent methods.) Thus in section 2, one can think of GMRES as a *generic* minimal residual method that characterizes corrections by (1.2). In section 3, we discuss further how singularity or ill conditioning can appear in GMRES and affect its practical performance. We outline an efficient and reliable way of detecting singularity or ill conditioning when it threatens to cause breakdown or otherwise degrade the performance of the method. In section 4, we discuss several illustrative numerical experiments.

Others have considered GMRES and related methods on singular or ill-conditioned systems. It is noted in [3] and [15] that GMRES can be used to solve singular homogeneous systems that arise in Markov chain modeling. In [9], conditions are given for the convergence of general Krylov subspace methods on singular systems, and particular results are derived for the QMR [10] and TFQMR [7] methods (see section 2 below), with applications to Markov chain modeling. Deflation-like modifications of GMRES based on truncated singular value decomposition solutions have recently been considered in [12]; see also [13] and the references in [12] and [13] for more on deflation techniques for nearly singular systems. In [14], extensions of GMRES are considered in which Krylov subspaces are augmented with approximate eigenvectors generated during previous iterations. These extensions appear to be most effective when there are a few relatively small eigenvalues.

In the following, we denote the null space and range of A by $\mathcal{N}(A)$ and $\mathcal{R}(A)$, respectively, and say that (1.1) is *consistent* if $b \in \mathcal{R}(A)$. We set $r_k \equiv b - Ax_k$ for each k and denote the restriction of A to a subspace $\mathcal{S} \subseteq \mathbb{R}^n$ by $A|_{\mathcal{S}}$. As a convention, we always regard x_0 as determined without breakdown at the "0th" step and define $\mathcal{K}_0 \equiv \{0\}$. Also, we assume that GMRES terminates immediately upon breakdown.

2. Theoretical discussion. Although our interest is primarily in (1.1) when A is singular, the results in this section also apply, as appropriate, when A is nonsingular. The questions of interest are the following:

- Will GMRES determine a least-squares solution without breakdown?
- When has a least-squares solution been reached by GMRES?
- When is a least-squares solution determined by GMRES the pseudoinverse solution?
- How ill conditioned can the GMRES least-squares problem (1.2) become?

We begin with several general results.

LEMMA 2.1. *Apply GMRES to* (1.1) *and suppose that* $\dim \mathcal{K}_k = k$ *for some* $k \geq 0$. *Then exactly one of the following holds:*

(i) $\dim A(\mathcal{K}_k) = k - 1$ *and* $A(x_0 + z) \neq b$ *for every* $z \in \mathcal{K}_k$;
(ii) $\dim A(\mathcal{K}_k) = k$, $\dim \mathcal{K}_{k+1} = k$, x_k *is uniquely defined, and* $Ax_k = b$;
(iii) $\dim A(\mathcal{K}_k) = k$, $\dim \mathcal{K}_{k+1} = k + 1$, x_k *is uniquely defined, and* $Ax_k \neq b$.

Proof. First, note that if $\dim \mathcal{K}_k = k$ and $k > 0$, then $\dim A(\mathcal{K}_{k-1}) = k - 1$. Indeed, in this case r_0, Ar_0, ..., $A^{k-1}r_0$ constitute a basis of $\mathcal{K}_k$ and, therefore,

$Ar_0, \ldots, A^{k-1}r_0$ constitute a basis of $A(\mathcal{K}_{k-1})$. With this observation and the fact that $A(\mathcal{K}_{k-1}) \subseteq A(\mathcal{K}_k)$ for $k > 0$, it is clear that the assumption $\dim \mathcal{K}_k = k$ implies $k - 1 \leq \dim A(\mathcal{K}_k) \leq k$ for all $k \geq 0$. Note also that $r_0 \notin A(\mathcal{K}_{k-1})$ if $k > 0$.

If $\dim A(\mathcal{K}_k) = k - 1$, then conclusions (ii) and (iii) cannot hold. Furthermore, $k > 0$ and $A(\mathcal{K}_{k-1}) = A(\mathcal{K}_k)$ in this case, and, since $r_0 \notin A(\mathcal{K}_{k-1})$, it follows that $r_0 \notin A(\mathcal{K}_k)$. Then $A(x_0 + z) \neq b$ for every $z \in \mathcal{K}_k$, and (only) conclusion (i) holds.

Suppose that $\dim A(\mathcal{K}_k) = k$. Then x_k is uniquely defined; furthermore, since $A(\mathcal{K}_k) \subseteq \mathcal{K}_{k+1}$, we have $k = \dim A(\mathcal{K}_k) \leq \dim \mathcal{K}_{k+1} \leq k + 1$. If $\dim \mathcal{K}_{k+1} = k$, then we must have $A(\mathcal{K}_k) = \mathcal{K}_{k+1}$ and, hence, $r_0 \in A(\mathcal{K}_k)$. It follows from (1.2) that $r_k = 0$ and $Ax_k = b$; thus (only) conclusion (ii) holds. If $\dim \mathcal{K}_{k+1} = k + 1$, then $r_0 \notin A(\mathcal{K}_k)$, $r_k \neq 0$, $Ax_k \neq b$, and (only) conclusion (iii) holds. ∎

This lemma implies the following result.

THEOREM 2.2. *Apply GMRES to* (1.1). *Then, at some step, either*

(a) *GMRES breaks down through rank deficiency of the least-squares problem* (1.2) *without determining a solution or*

(b) *GMRES determines a solution without breakdown and then breaks down at the next step through degeneracy of the Krylov subspace.*

Proof. We have $\dim \mathcal{K}_0 = 0$. Assume that for some $k \geq 0$ GMRES has proceeded to the kth step with $\dim \mathcal{K}_k = k$. Then exactly one of the three conclusions of Lemma 2.1 must hold. If conclusion (i) holds, then we have (a) above. If conclusion (ii) holds, then we have (b). If conclusion (iii) holds, then $Ax_k \neq b$ and the iteration continues to the next step. The theorem follows by induction. ∎

The alternatives of this theorem give useful insights into the eventual outcome of applying GMRES to (1.1). For example, if (1.1) is not consistent, then breakdown through rank deficiency of (1.2) will eventually occur; in practice, this may be preceded by dangerous ill conditioning, as discussed further below. Conversely, breakdown through degeneracy of the Krylov subspace occurs if and only if (1.1) is consistent and the solution has been found. Also, these results imply the result in [16, Prop. 2, p. 865] cited earlier: if A is nonsingular, then GMRES does not break down until the solution of (1.1) has been found. Indeed, if A is nonsingular, then GMRES cannot break down through rank deficiency of (1.2), and the second alternative must hold. However, the reader is cautioned to make inferences carefully; e.g., Example 1.1 above shows that there can be breakdown through rank deficiency in the consistent case before a solution is found.

The next result characterizes circumstances in which a least-squares solution has been reached.

LEMMA 2.3. *At the kth step, GMRES determines a least-squares solution of* (1.1) *without breakdown if and only if*

$$\dim A^T(\mathcal{K}_{k+1}) = \dim A(\mathcal{K}_k) = k. \tag{2.1}$$

Proof. By definition, GMRES does not break down at the kth step if and only if $\dim A(\mathcal{K}_k) = k$. Thus we need only show that x_k is a least-squares solution of (1.1) if and only if $\dim A^T(\mathcal{K}_{k+1}) = \dim A(\mathcal{K}_k)$.

From (1.2), we have that x_k is a least-squares solution of (1.1) if and only if it is possible to reach a least-squares solution of (1.1) through *some* correction in $\mathcal{K}_k$, i.e., if and only if there is some $z \in \mathcal{K}_k$ such that

$$0 = A^T [b - A(x_0 + z)] = A^T(r_0 - Az). \tag{2.2}$$

But (2.2) holds for some $z \in \mathcal{K}_k$ if and only if $A^T r_0 \in A^T A(\mathcal{K}_k)$, which is equivalent to $A^T(\mathcal{K}_{k+1}) = A^T A(\mathcal{K}_k)$. To complete the proof, we note that $\dim A^T A(\mathcal{K}_k) =$

$\dim A(\mathcal{K}_k)$. Indeed, we clearly have $\dim A^T A(\mathcal{K}_k) \leq \dim A(\mathcal{K}_k)$. If $\dim A^T A(\mathcal{K}_k) < \dim A(\mathcal{K}_k)$, then there is a $w \in \mathcal{K}_k$ such that $Aw \neq 0$ and $A^T Aw = 0$. But then $0 = w^T A^T Aw = \|Aw\|_2^2$, which is a contradiction. ∎

With Lemma 2.1, one can easily extend Lemma 2.3 to conclude additionally that if (2.1) holds, then (1.1) is consistent if and only if $\dim \mathcal{K}_{k+1} = k$; i.e., GMRES breaks down at step $k+1$ through degeneracy of the Krylov subspace.

We use Lemma 2.3 to characterize the property of A that yields the most satisfactory answers to the questions posed at the beginning of this section. This property is $\mathcal{N}(A) = \mathcal{N}(A^T)$, equivalently, $\mathcal{N}(A) = \mathcal{R}(A)^\perp$, which holds when A is normal, e.g., when it is symmetric or skew symmetric. It also clearly holds when A is nonsingular. In general, this property holds if and only if $\mathcal{N}(A)^\perp$ is an invariant subspace of A. Also, it holds only if all eigenvectors of A associated with nonzero eigenvalues are orthogonal to $\mathcal{N}(A)$. Note that it does *not* hold for the matrices of Example 1.1; indeed, it holds for $A \in \mathbb{R}^{2\times 2}$ if and only if A is either nonsingular or symmetric. Neither does it hold for the "shift" operator of Example 1.2.

THEOREM 2.4. *GMRES determines a least-squares solution of* (1.1) *without breakdown for all* b *and* x_0 *if and only if* $\mathcal{N}(A) = \mathcal{N}(A^T)$. *If* $\mathcal{N}(A) = \mathcal{N}(A^T)$ *and a least-squares solution is reached at some step, then GMRES breaks down at the next step, with breakdown through degeneracy of the Krylov subspace if* (1.1) *is consistent and through rank deficiency of the least-squares problem* (1.2) *otherwise. Furthermore, if* (1.1) *is consistent and* $x_0 \in \mathcal{R}(A)$, *then the solution reached is the pseudoinverse solution.*

Proof. First, suppose that $\mathcal{N}(A) \neq \mathcal{N}(A^T)$. One can choose b and x_0 such that $r_0 \in \mathcal{N}(A)$ and $A^T r_0 \neq 0$. Then x_0 is not a least-squares solution. Furthermore, $\dim A(\mathcal{K}_1) = 0$, so GMRES breaks down at the first step before reaching a least-squares solution.

Now assume $\mathcal{N}(A) = \mathcal{N}(A^T)$. Then for each k, we have $\dim A^T(\mathcal{K}_{k+1}) = \dim A(\mathcal{K}_{k+1})$, and (2.1) becomes

$$\dim A(\mathcal{K}_{k+1}) = \dim A(\mathcal{K}_k) = k.$$

This condition must hold for some k, $0 \leq k \leq n$, and it follows from Lemma 2.3 that GMRES determines a least-squares solution x_k without breakdown at the kth step. Furthermore, since $\dim A(\mathcal{K}_{k+1}) = k$, GMRES breaks down at step $k+1$. One concludes from Theorem 2.2 that breakdown is through degeneracy of the Krylov subspace if (1.1) is consistent and through rank deficiency of the least-squares problem (1.2) otherwise. If (1.1) is consistent, then x_k is a solution and, furthermore, $\mathcal{K}_k \subseteq \mathcal{R}(A)$. If in addition $x_0 \in \mathcal{R}(A)$, then $x_k = x_0 + z_k \in x_0 + \mathcal{K}_k \subseteq \mathcal{R}(A) = \mathcal{N}(A)^\perp$. Since a (least-squares) solution of (1.1) is the pseudoinverse solution if and only if it lies in $\mathcal{N}(A)^\perp$, it follows that x_k is the pseudoinverse solution. ∎

If it is known that $\mathcal{N}(A) = \mathcal{N}(A^T)$, then Theorem 2.4 provides theoretical assurance not only that GMRES will determine a least-squares solution of (1.1) without breakdown but also that reaching it will be indicated by breakdown at the next step. If (1.1) is consistent as well, then choosing $x_0 \in \mathcal{R}(A)$, e.g., $x_0 = 0$, will yield the pseudoinverse solution without breakdown, and reaching it will be indicated by zero residual norm.

If $\mathcal{N}(A) = \mathcal{N}(A^T)$ and (1.1) is consistent, then the least-squares problem (1.2) will remain as well conditioned as the nature of A will allow until a solution of (1.1) is reached. Indeed, if we denote

$$A_k \equiv A|_{\mathcal{K}_k},$$

then the appropriate condition number for (1.2) is $\kappa_2(A_k)$, which satisfies

$$\kappa_2(A_k) \equiv \frac{\max\limits_{z\in\mathcal{K}_k,\ z\neq 0} \|Az\|_2 \big/ \|z\|_2}{\min\limits_{z\in\mathcal{K}_k,\ z\neq 0} \|Az\|_2 \big/ \|z\|_2} \leq \frac{\max\limits_{z\in\mathcal{R}(A),\ z\neq 0} \|Az\|_2 \big/ \|z\|_2}{\min\limits_{z\in\mathcal{R}(A),\ z\neq 0} \|Az\|_2 \big/ \|z\|_2} \equiv \kappa_2(A|_{\mathcal{R}(A)}) \tag{2.3}$$

since $\mathcal{K}_k \subseteq \mathcal{R}(A)$ in the consistent case. Note that, since $\mathcal{R}(A) = \mathcal{N}(A^T)^\perp = \mathcal{N}(A)^\perp$, $\kappa_2(A|_{\mathcal{R}(A)})$ is just the ratio of the largest singular value of A to the smallest positive one. Also, recall from above that, in the consistent case, if a solution is reached at some step, then breakdown of GMRES at the next step occurs because of degeneracy of the Krylov subspace and not because of rank deficiency of the least-squares problem (1.2). These reassuring results are to be expected, for if $\mathcal{N}(A) = \mathcal{N}(A^T)$ and (1.1) is consistent, then everything reduces to the nonsingular case on $\mathcal{R}(A) = \mathcal{N}(A)^\perp$.

If $\mathcal{N}(A) = \mathcal{N}(A^T)$ but (1.1) is not consistent, then, despite the theoretical guarantee of Theorem 2.4 that GMRES will not break down, the least-squares problem (1.2) may necessarily become dangerously ill conditioned before a least-squares solution of (1.1) is reached, regardless of the conditioning of $A|_{\mathcal{R}(A)}$. This is shown by Theorem 2.5 below. It is, perhaps, not surprising, because if a least-squares solution is reached at some step, then, in the inconsistent case, breakdown at the next step occurs because of rank deficiency of the least-squares problem (1.2), rather than degeneracy of the Krylov subspace.

THEOREM 2.5. *Suppose that $\mathcal{N}(A) = \mathcal{N}(A^T)$, and denote the least-squares residual for* (1.1) *by r_*. If $r_{k-1} \neq r_*$ for some k, then*

$$\kappa_2(A_k) \geq \frac{\|A_k\|_2}{\|\bar{A}_k\|_2} \cdot \frac{\|r_{k-1}\|_2}{\sqrt{\|r_{k-1}\|_2^2 - \|r_*\|_2^2}}, \tag{2.4}$$

where $A_k \equiv A|_{\mathcal{K}_k}$ and $\bar{A}_k \equiv A|_{\mathcal{K}_k + \operatorname{span}\{r_*\}}$.

Proof. Note that $r_* \in \mathcal{R}(A)^\perp = \mathcal{N}(A)$ and $r_{k-1} - r_* \in \mathcal{R}(A) = \mathcal{N}(A)^\perp$. Then, since $r_{k-1} - r_* \in \mathcal{K}_k + \operatorname{span}\{r_*\}$, we have

$$\begin{aligned} \|Ar_{k-1}\|_2 &= \|A(r_{k-1} - r_* + r_*)\|_2 = \|A(r_{k-1} - r_*)\|_2 \\ &\leq \|\bar{A}_k\|_2 \cdot \|r_{k-1} - r_*\|_2 = \|\bar{A}_k\|_2 \cdot \sqrt{\|r_{k-1}\|_2^2 - \|r_*\|_2^2}, \end{aligned}$$

whence

$$\frac{\|Ar_{k-1}\|_2}{\|r_{k-1}\|_2} \leq \|\bar{A}_k\|_2 \cdot \frac{\sqrt{\|r_{k-1}\|_2^2 - \|r_*\|_2^2}}{\|r_{k-1}\|_2}. \tag{2.5}$$

Since $r_{k-1} \in \mathcal{K}_k$, (2.4) follows from (2.5) and the definition of $\kappa_2(A_k)$ (see (2.3)). ∎

If (1.1) is consistent, then $r_* = 0$ and $\bar{A}_k = A_k$. It follows that (2.4) is just the trivial bound $\kappa_2(A_k) \geq 1$ in this case. In general, we have $1 \geq \|A_k\|_2/\|\bar{A}_k\|_2 \geq \|A_k\|_2/\|A\|_2$, and (2.4) yields

$$\kappa_2(A_k) \geq \frac{\|A_k\|_2}{\|A\|_2} \cdot \frac{\|r_{k-1}\|_2}{\sqrt{\|r_{k-1}\|_2^2 - \|r_*\|_2^2}}, \tag{2.6}$$

which may be more easily applied in the inconsistent case.

If A is singular and $\mathcal{N}(A) = \mathcal{N}(A^T)$, then it is evident from (2.6) that, for an unfortunate choice of b and x_0, the least-squares problem (1.2) will become so ill

conditioned before breakdown that little or no accuracy can be expected in a solution computed in finite-precision arithmetic. Indeed, in view of (2.6), one would expect that, in many cases, the residual for the computed solution will first decrease in norm for a number of iterations and then lose accuracy and perhaps increase as a least-squares solution is approached and accuracy is degraded by increasing ill conditioning. (This is seen in Experiment 4.2 below.) In such cases, it would clearly be desirable to terminate the iterations when approximately optimal accuracy has been reached. Note that the usual termination criteria based on the size of the residual norm are unlikely to be of any use in this case; some alternative criterion is needed.

We show how (2.6) can be used to derive a heuristic guideline for terminating the iterations at an approximately optimal point in finite-precision arithmetic. We make two assumptions that are reasonable but by no means the only possible assumptions; our main purpose is to demonstrate the method of derivation. (The guideline resulting from these assumptions is borne out well in Experiment 4.2 below.) The first assumption is that $\kappa_2(A_k)$ is about as small as possible, given the lower bound (2.6), i.e., that

$$\kappa_2(A_k) \approx \frac{\|A_k\|_2}{\|A\|_2} \cdot \frac{\|r_{k-1}\|_2}{\sqrt{\|r_{k-1}\|_2^2 - \|r_*\|_2^2}}.$$

The second assumption is that the computed value of r_k, denoted by $\hat{r}_k$, satisfies

$$\frac{\|\hat{r}_k - r_k\|_2}{\|r_0\|_2} \approx \mathbf{u}\kappa_2(A_k),$$

where $\mathbf{u}$ is unit rounding error. A rigorous worst-case bound on $\|\hat{r}_k - r_k\|_2/\|r_0\|_2$ would require $\mathbf{u}\kappa_2(A_k)$ multiplied by a polynomial of low degree in n and k (see [11, Chap. 5]), but this is not necessary here. With these assumptions, we have

$$\begin{aligned}
\frac{\|\hat{r}_k - r_*\|_2}{\|r_0\|_2} &\leq \frac{\|\hat{r}_k - r_k\|_2}{\|r_0\|_2} + \frac{\|r_k - r_*\|_2}{\|r_0\|_2} \\
&\approx \mathbf{u}\kappa_2(A_k) + \frac{\sqrt{\|r_k\|_2^2 - \|r_*\|_2^2}}{\|r_0\|_2} \\
&\leq \mathbf{u}\kappa_2(A_k) + \frac{\sqrt{\|r_{k-1}\|_2^2 - \|r_*\|_2^2}}{\|r_0\|_2} \\
&\approx \mathbf{u}\kappa_2(A_k) + \frac{\|A_k\|_2}{\|A\|_2} \cdot \frac{\|r_{k-1}\|_2}{\|r_0\|_2} \cdot \frac{1}{\kappa_2(A_k)} \\
&= B(\kappa_2(A_k)),
\end{aligned} \tag{2.7}$$

where

$$B(\kappa) \equiv \mathbf{u}\kappa + \frac{\|A_k\|_2}{\|A\|_2} \cdot \frac{\|r_{k-1}\|_2}{\|r_0\|_2} \cdot \frac{1}{\kappa}.$$

It is easily seen that B is minimized when

$$\kappa = \kappa_{\min} \equiv \sqrt{\frac{\|A_k\|_2}{\|A\|_2} \cdot \frac{\|r_{k-1}\|_2}{\|r_0\|_2} \cdot \frac{1}{\mathbf{u}}}, \tag{2.8}$$

which suggests a heuristic guideline as follows: If the iterations are terminated with $\kappa_2(A_k) \approx \kappa_{\min}$ given by (2.8), then (2.7) gives an approximate minimal bound

$$\frac{\|\hat{r}_k - r_*\|_2}{\|r_0\|_2} \leq B(\kappa_{\min}) = 2\sqrt{\frac{\|A_k\|_2}{\|A\|_2} \cdot \frac{\|r_{k-1}\|_2}{\|r_0\|_2} \cdot \mathbf{u}}\,. \tag{2.9}$$

This can be simplified for practical purposes by assuming that $\|A_k\|_2/\|A\|_2 \approx 1$ and $\|r_{k-1}\|_2 \approx \|\hat{r}_{k-1}\|_2$. We discuss how to monitor $\kappa_2(A_k)$ efficiently in practice in section 3.

If $\mathcal{N}(A) \neq \mathcal{N}(A^T)$, then it follows from Theorem 2.4 that, for *some* b and x_0, GMRES will break down before determining a least-squares solution of (1.1). However, there is an important special case in which GMRES still reliably determines a least-squares solution, viz., that in which $\mathcal{N}(A) \cap \mathcal{R}(A) = \{0\}$ and (1.1) are consistent. This occurs, e.g., in Experiment 4.3 below.

THEOREM 2.6. *Suppose that $\mathcal{N}(A) \cap \mathcal{R}(A) = \{0\}$. If (1.1) is consistent, then GMRES determines a solution without breakdown at some step and breaks down at the next step through degeneracy of the Krylov subspace.*

Proof. Since (1.1) is consistent, $r_0 \in \mathcal{R}(A)$ and $\mathcal{K}_k \subseteq \mathcal{R}(A)$ for each k. Since $\mathcal{N}(A) \cap \mathcal{R}(A) = \{0\}$, this implies that $\dim A(\mathcal{K}_k) = \dim \mathcal{K}_k$ for each k. Then there cannot be breakdown through rank deficiency of the least-squares problem (1.2), and the theorem follows from Theorem 2.2. □

Conditions that are essentially equivalent to those in Theorem 2.6 appear in [9]. The *index* of A, denoted index(A), is defined to be the smallest integer q such that A^q and A^{q+1} have the same rank. It is easily seen that index$(A) = 1$ if and only if A is singular and $\mathcal{N}(A) \cap \mathcal{R}(A) = \{0\}$. For a consistent system (1.1) with index$(A) = 1$, general conditions are given in [9] under which a Krylov subspace method is convergent. It is further shown in [9] that the QMR and TFQMR methods are convergent for such a system.

If $\mathcal{N}(A) \cap \mathcal{R}(A) = \{0\}$ and (1.1) is consistent, then $\kappa_2(A_k)$ satisfies (2.3). However, note that if $\mathcal{N}(A) \neq \mathcal{N}(A^T)$, then $\min_{z \in \mathcal{R}(A),\ z \neq 0} \|Az\|_2/\|z\|_2$ may be smaller than the smallest positive singular value of A, and so $\kappa_2(A|_{\mathcal{R}(A)})$ may be larger than the ratio of the largest singular value of A to the smallest positive one. Still, the least-squares problem (1.2) is as well conditioned as the nature of A will allow and cannot become arbitrarily ill conditioned before a solution is determined by GMRES through an unfortunate choice of b and x_0. This is not surprising, since GMRES breakdown occurs because of degeneracy of the Krylov subspace, rather than rank deficiency of the least-squares problem (1.2). As when (1.1) is consistent and $\mathcal{N}(A) = \mathcal{N}(A^T)$, the setting reduces to the nonsingular case on $\mathcal{R}(A)$, although now $\mathcal{R}(A)$ may not be $\mathcal{N}(A)^\perp$. When (1.1) is not consistent, breakdown must occur because of rank deficiency of (1.2), and in general we cannot expect (1.2) to remain well conditioned, whether or not a least-squares solution is reached.

We conclude this section by noting that, in some applications, one can easily project b onto $\mathcal{R}(A)$. For example, in each of Experiments 4.2 and 4.3 below, $\mathcal{N}(A^T)$ is one dimensional, and it is not difficult to determine a unit vector in $\mathcal{N}(A^T)$ and then to project b onto $\mathcal{N}(A^T)^\perp = \mathcal{R}(A)$. In such an application, if GMRES can be expected to behave well on a consistent system, e.g., if $\mathcal{N}(A) = \mathcal{N}(A^T)$ or $\mathcal{N}(A) \cap \mathcal{R}(A) = \{0\}$, then it is clearly desirable to project b onto $\mathcal{R}(A)$ before starting GMRES. By doing this, one can determine a least-squares solution for the original b without risking the dangerous ill conditioning that may precede GMRES breakdown with rank deficiency

of (1.2). In addition, if $\mathcal{N}(A) = \mathcal{N}(A^T)$, then one can determine the pseudoinverse solution by taking $x_0 \in \mathcal{R}(A)$, e.g., $x_0 = 0$.

3. Practical handling of (near) singularity. In section 2, we considered the conditioning of the least-squares problem (1.2) and how it might be affected by A and perhaps b and x_0. In this section, we look further into how singularity or ill conditioning can arise in GMRES and discuss how conditioning can be monitored efficiently in practice.

Recall from section 1 that, prior to breakdown, an implementation of GMRES generates a basis of $\mathcal{K}_k$ for each k. We denote the matrix having the basis vectors as columns by $B_k \in \mathbb{R}^{n\times k}$. The kth GMRES correction z_k, which is the solution of (1.2), is not computed for each k, but when desired, it is determined by first finding y_k that solves

$$\min_{y\in\mathbb{R}^k} \|r_0 - AB_k y\|_2 \tag{3.1}$$

and then forming $z_k = B_k y_k$. Thus ill conditioning or singularity is a concern in GMRES only if it becomes manifested in ill conditioning or rank deficiency of AB_k or B_k.

Sound GMRES implementations are designed so that, as much as possible, each B_k is well conditioned regardless of the conditioning of A. For example, the standard implementation of [16] and Householder variants in [18] determine ideally conditioned B_k such that $B_k^T B_k = I_k$ (in exact arithmetic). Other implementations in [2] and [19] generate B_k that are usually well conditioned, if not ideally conditioned. In any event, in well-constructed GMRES implementations, the conditioning of B_k does not suffer directly from ill conditioning of A; furthermore, any ill conditioning of B_k seems likely to be reflected in ill conditioning of AB_k. Therefore, we focus on the conditioning of AB_k here.

In practice, a reasonable course is to monitor the conditioning of AB_k and terminate the GMRES iterations if excessive ill conditioning or rank deficiency appears. Typically, the solution of (3.1) is computed using a factorization $AB_k = Q_k R_k$, where $Q_k \in \mathbb{R}^{n\times k}$ has orthonormal columns and $R_k \in \mathbb{R}^{k\times k}$ is upper triangular. It is reasonable to assume that this factorization is determined using one or more stable factorization techniques. For example, the implementations of [16] and [18] first use modified Gram–Schmidt or, respectively, Householder transformations to produce $AB_k = B_{k+1}H_k$, where $H_k \in \mathbb{R}^{(k+1)\times k}$ is upper Hessenberg, and then use plane rotations $J_1, \ldots, J_k$ to obtain $A_k B_k = Q_k R_k$ with $Q_k = B_{k+1}J_1^T \ldots J_k^T (I_k, 0)^T$ and $R_k = (I_k, 0)J_k \ldots J_1 H_k$. In general, each Q_k may be only implicitly specified, as in the implementations of [16] and [18], but each R_k is always produced explicitly. Then, since the conditioning of AB_k is determined by that of R_k, it suffices to monitor the conditioning of R_k and terminate the iterations if excessive ill conditioning or singularity appears.

In the important case in which $B_k^T B_k = I_k$, as in the implementations of [16] and [18], we have $\kappa_2(R_k) = \kappa_2(AB_k) = \kappa_2(A_k) \le \kappa_2(A)$, where $A_k = A|_{\mathcal{K}_k}$ as above. This inequality need not be strict; for example, if A is nonsingular and GMRES proceeds for n steps without breakdown, then $A_n = A$ and $\kappa_2(R_n) = \kappa_2(A_n) = \kappa_2(A)$. Thus R_k can become fully as ill conditioned as A. However, if r_0 lies in an invariant proper subspace, then $\kappa_2(R_k)$ may remain much less than $\kappa_2(A)$. The following example illustrates extreme behavior.

Example 3.1. Assume that $B_k^T B_k = I_k$ for each k. Suppose that we have $\sigma_1 \ge$

$\cdots \geq \sigma_{n-1} = \sigma_n > 0$, and define

$$A \equiv \begin{pmatrix} 0 & \cdots & 0 & \sigma_{n-1} & 0 \\ \sigma_1 & \ddots & \vdots & 0 & 0 \\ 0 & \ddots & 0 & \vdots & \vdots \\ \vdots & \ddots & \sigma_{n-2} & 0 & 0 \\ 0 & \cdots & 0 & 0 & \sigma_n \end{pmatrix}.$$

Clearly, $\sigma_1, \ldots, \sigma_n$ are the singular values of A, and $\kappa_2(A) = \sigma_1/\sigma_n$. For $i = 1, \ldots, n$, let e_i denote the ith column of I_n. If $r_0 = e_1$, then we have $\mathcal{K}_k = \text{span}\{e_1, \ldots, e_k\}$ and $\kappa_2(R_k) = \kappa_2(A_k) = \sigma_1/\sigma_k$ for $k = 1, \ldots, n-1$. In particular, the solution is reached at the $(n-1)$st step with $\kappa_2(R_{n-1}) = \sigma_1/\sigma_{n-1} = \sigma_1/\sigma_n = \kappa_2(A)$. However, if $r_0 = e_n$, then the solution is reached at the first step with $\kappa_2(R_1) = \sigma_n/\sigma_n = 1$.

A very efficient means of monitoring the conditioning of R_k is provided by *incremental condition estimation* (ICE) [4], [5]. This determines estimates of the largest and smallest singular values of each R_k in $O(k)$ arithmetic operations, given estimates of the largest and smallest singular values of R_{k-1}. Thus one can begin with $k = 1$ and use ICE to estimate incrementally the condition number of each successive R_k as k increases. Over a cycle of m GMRES steps, the total cost of estimating the condition number of each R_k, $1 \leq k \leq m$, is $O(m^2)$ arithmetic operations, which is negligible in most applications. A well-developed Fortran implementation of ICE is provided by auxiliary routine xLAIC1 of LAPACK [1], where x = S for single precision or x = D for double precision. This implementation was used in all of the numerical experiments reported in section 4.

4. Numerical experiments. In this section, we discuss several numerical experiments that illustrate the theoretical and practical points brought out above. A standard modified Gram–Schmidt GMRES implementation, as originally outlined in [16], was used in all experiments. Recall that with this implementation, the basis matrix B_k is ideally conditioned, with $B_k^T B_k = I_k$. This implementation was augmented with routine DLAIC1 of LAPACK for monitoring conditioning of the triangular factor of AB_k as discussed above. In all experiments, we took the zero vector to be the initial approximate solution and specified a stopping tolerance tol so that the GMRES iterations would terminate when $\|r_k\|_2 \leq \text{tol}\|b\|_2$. Of course, there was no expectation of stopping on the basis of such a test in cases in which (1.1) was not consistent; in these cases, termination was based on other criteria noted below. All computing was done in double precision Fortran on Sun Microsystems Sparc architectures.

Experiment 4.1. This experiment, which involves a contrived problem, points out the danger of not monitoring the conditioning of AB_k and terminating when excessive ill conditioning appears. The matrix A is from the example in [6, sect. 6],

$$A = \begin{pmatrix} 0 & 1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & -1 & 0 \end{pmatrix}.$$

We assume that n is odd, in which case A is singular with

$$\mathcal{N}(A) = \text{span}\{(1, 0, 1, 0, \ldots, 0, 1)^T\}.$$

Since A is skew symmetric, the conclusions of Theorem 2.4 hold, at least in exact arithmetic, and GMRES should find a least-squares solution of (1.1) without breakdown and then exhibit breakdown at the next step. In floating point arithmetic, however, GMRES produced misleading results.

We took $n = 49$, tol $= 10^{-6}$ and first ran GMRES with

$$b = (1/\sqrt{2}, 0, \ldots, 0, -1/\sqrt{2})^T,$$

for which (1.1) is consistent. GMRES safely terminated with a computed residual norm of 1.57×10^{-16} when the pseudoinverse solution was reached at the 24th step; the largest observed condition number estimate was 12.7. We then ran GMRES with $b = (1/\sqrt{2}, 0, \ldots, 0, 1/\sqrt{2})^T$, for which (1.1) is not consistent; the least-squares residual norm is $\sqrt{2}/5$. In exact arithmetic, a least-squares solution would have been obtained at the 24th step, and this would have been indicated by breakdown at the 25th step in the form of rank deficiency in the least-squares problems (1.2) and (3.1). Because of rounding error, exact breakdown did not occur, nor were any arithmetic exceptions such as overflow observed. However, the condition number estimate went from 12.7 at the 24th step to 1.47×10^{16} and 1.79×10^{30} at the 25th and 26th steps, respectively. We allowed GMRES to continue, restarting every 49 steps, until it declared *successful* termination at the 185th step with a computed residual norm of 6.68×10^{-7}. Of course, this was the value of the residual norm maintained recursively by GMRES and not the true residual norm, which was 9.14×10^{12} on termination!

We also note that the GMRES implementation used in these experiments did not re-evaluate the residual and its norm directly at each restart; i.e., it did not multiply the current approximate solution by A and subtract the result from b. Instead, it updated the residual at each restart by forming a certain linear combination of the Arnoldi basis vectors generated in the previous cycle of steps. Such updating saves an A-product at each restart and is usually a safe thing to do, unless extreme residual norm reduction is desired. In this example, however, it was not safe, and re-evaluating the residual directly at restarts would have indicated that GMRES had gone astray.

The next two experiments involve discretizations of boundary value problems for the partial differential equation

$$\Delta u + d\frac{\partial u}{\partial x_1} = f(x), \quad x = (x_1, x_2) \in \Omega \equiv [0,1] \times [0,1], \tag{4.1}$$

where d is a constant and f is a given function. In the experiments reported here, we discretized (4.1) with the usual second-order centered differences on a 100×100 mesh of equally spaced discretization points, so that the resulting linear systems were of dimension $10,000$. We took $d = 10$ and preconditioned the discretized problems on the right with a fast Poisson solver from FISHPACK [17]. This preconditioner is very effective for this fairly small value of d. We took tol $= 10^{-10}$ in order to see how GMRES behaved with a tight stopping tolerance. We also stopped the iterations when the condition number estimate became greater than $1/(50\mathbf{u}) \approx 10^{14}$. In the trials outlined below, there was no need to restart GMRES; in each case, there was termination because of either sufficient residual norm reduction or excessive ill conditioning before the maximum allowable number of iterations (50) had been reached.

In each of these two experiments, it is possible to give a simple characterization of $\mathcal{N}(A^T)$. In each, then, we first consider a b for which (1.1) is not consistent and then project it onto $\mathcal{R}(A)$ to obtain a consistent system that is effectively solved by

TABLE 4.1
GMRES iterations 9–19 on problem (4.1) with periodic boundary conditions.

Iteration no.	GMRES recursive residual norm	Computed residual norm	Condition no. estimate
9	99.000000080681	99.000000080680	7.80×10^3
10	99.000000005202	99.000000005201	4.17×10^4
11	99.000000000146	99.000000000145	1.65×10^5
12	99.000000000008	99.000000000007	9.97×10^5
13	99.000000000002	99.000000000000	4.71×10^6
14	99.000000000002	99.000000000000	3.20×10^7
15	99.000000000001	99.000000000001	1.76×10^8
16	98.999999999935	99.000000000068	1.33×10^9
17	98.999999997323	99.000000002679	8.41×10^9
18	98.999999811806	99.000000188196	7.05×10^{10}
19	98.999990468226	99.000009534599	5.02×10^{11}

GMRES. The result is both an approximate solution of the consistent system and an approximate least-squares solution of the original inconsistent system.

Experiment 4.2. In this experiment, we imposed periodic boundary conditions: $u(x_1, 0) = u(x_1, 1)$ and $u(0, x_2) = u(1, x_2)$ for $0 \le x_1, x_2 \le 1$. The matrix A is given as follows:

$$A = \frac{1}{h^2}\begin{pmatrix} T_m & I_m & & I_m \\ I_m & \ddots & \ddots & \\ & \ddots & \ddots & I_m \\ I_m & & I_m & T_m \end{pmatrix}, \quad T_m = \begin{pmatrix} -4 & \alpha_+ & & \alpha_- \\ \alpha_- & \ddots & \ddots & \\ & \ddots & \ddots & \alpha_+ \\ \alpha_+ & & \alpha_- & -4 \end{pmatrix} \in \mathbb{R}^{m\times m},$$

and $m = \sqrt{n} = 100$, $h = 1/m$, and $\alpha_\pm = 1 \pm dh/2$. It is easy to verify that A is normal and that

$$\mathcal{N}(A) = \mathcal{N}(A^T) = \text{span}\{(1, 1, \ldots, 1)^T\}; \tag{4.2}$$

then Theorems 2.4 and 2.5 are applicable.

We first took b to be a discretization of $f(x) = x_1 + x_2$. For this b, (1.1) is not consistent; the least-squares residual norm is 99. GMRES began with an initial residual norm of 107.1 and terminated after 21 iterations with a condition number estimate greater than the termination value $1/(50\mathbf{u}) \approx 10^{14}$. A subset of the iterations is shown in Table 4.1, which gives to 14-digit accuracy both the residual norm values maintained recursively by GMRES and the directly computed residual norms, as well as the condition number estimates. Note that the two norm values agree well and decrease toward the least-squares residual norm through iteration 15, but then the computed norms begin to increase while the recursive norm values continue erroneously to decrease below the least-squares residual norm. Since $\mathbf{u} \approx 2.2 \times 10^{-16}$ here, the heuristic guideline developed in section 2 would have called for termination when the condition number estimate was about 10^8. Table 4.1 shows that this would have been a very good point at which to terminate: the computed residual norm would have been near its minimum value, and the recursive residual norm value would have still been accurate. Note the pessimism of the bound (2.9) in this case.

Using the characterization of $\mathcal{N}(A^T)$ in (4.2), we next projected the above b onto $\mathcal{R}(A)$ to obtain a consistent system. The initial residual norm was 40.82. After 17 iterations, GMRES successfully met the termination test based on tol $= 10^{-10}$ and

terminated with a residual norm of 2.441×10^{-9}. No major inaccuracy was observed; the recursive residual norm value agreed with the directly computed residual norm to five significant digits. Since $\mathcal{N}(A) = \mathcal{N}(A^T)$ and the initial guess was zero, the final iterate was an approximate pseudoinverse solution of not only the consistent system but also the inconsistent system with the original b.

Experiment 4.3. In this experiment, we imposed Neumann boundary conditions: $\partial u(x)/\partial \nu = 0$ for $x \in \partial\Omega$. The matrix A is now given by

$$A = \frac{1}{h^2}\begin{pmatrix} T_m & 2I_m & & & \\ I_m & T_m & I_m & & \\ & \ddots & \ddots & \ddots & \\ & & I_m & T_m & I_m \\ & & & 2I_m & T_m \end{pmatrix}, \quad T_m = \begin{pmatrix} -4 & 2 & & & \\ \alpha_- & -4 & \alpha_+ & & \\ & \ddots & \ddots & \ddots & \\ & & \alpha_- & -4 & \alpha_+ \\ & & & 2 & -4 \end{pmatrix} \in \mathbb{R}^{m\times m},$$

and m, h, and $\alpha_\pm$ are as in Experiment 4.2. We have $\mathcal{N}(A) = \text{span}\{(1, 1, \cdots, 1)^T\}$ as before, but now $\mathcal{N}(A^T) \neq \mathcal{N}(A)$. Indeed, we determine $\mathcal{N}(A^T)$ as follows: Set

$$D_m \equiv \text{diag}\,(1, 2/\alpha_-, 2\alpha_+/\alpha_-^2, \ldots, 2\alpha_+^{m-3}/\alpha_-^{m-2}, \alpha_+^{m-2}/\alpha_-^{m-2}) \in \mathbb{R}^{m\times m},$$

and define a block-diagonal matrix $D = \text{diag}\,(D_m, 2D_m, \ldots, 2D_m, D_m) \in \mathbb{R}^{n\times n}$. Then one can verify that DA is symmetric, and it therefore follows that $\mathcal{N}(A^T) = \text{span}\{D(1, 1, \ldots, 1)^T\}$. With this characterization of $\mathcal{N}(A^T)$, one sees that $\mathcal{N}(A) \cap \mathcal{R}(A) = \{0\}$; then Theorem 2.6 applies when (1.1) is consistent.

The procedures and observations in this experiment were much like those in Experiment 4.2. We first took b to be a discretization of $f(x) = x_1 + x_2 + \sin 10x_1 \cos 10x_2 + e^{10x_1x_2}$. This gave somewhat more dramatic results than the choice of f in Experiment 4.2. For this b, (1.1) is not consistent; the least-squares residual is 5.302×10^4. GMRES began with an initial residual norm of 1.232×10^5 and terminated after 30 iterations with a condition number estimate greater than $1/(50\mathbf{u}) \approx 10^{14}$. The final computed residual norm was 6.305×10^4, which suggests that the GMRES iterates were not converging to a least-squares solution (at least not in any practical sense, given the very large condition number). We next used the characterization $\mathcal{N}(A^T) = \text{span}\{D(1, 1, \ldots, 1)^T\}$ to project this b onto $\mathcal{R}(A)$ and to obtain a consistent system. The initial residual norm was 1.112×10^5. After 23 iterations, GMRES successfully met the termination test based on tol $= 10^{-10}$ and terminated with a residual norm of 8.716×10^{-6}. No major inaccuracy was observed; the recursive residual norm agreed with the directly computed residual norm to three significant digits. In this case, the final iterate was not a pseudoinverse solution of either the consistent system or the inconsistent system with the original b.

5. Summary discussion. We have addressed the performance of GMRES on a linear system $Ax = b$ when A is singular or ill conditioned. Theoretical results are given that are of interest primarily in the singular case; these hold not only for GMRES but also for any mathematically equivalent method. In general, at some step, GMRES will either (a) break down through rank deficiency of the GMRES least-squares problem without determining a solution or (b) determine a solution without breakdown and then break down at the next step through degeneracy of the Krylov subspace.

More extensive results hold when $\mathcal{N}(A) = \mathcal{N}(A^T)$. This condition is necessary and sufficient for GMRES to determine a least-squares solution without breakdown for *all* b and x_0. If $\mathcal{N}(A) = \mathcal{N}(A^T)$ and the system is consistent, then the condition

number of the GMRES least-squares problem remains bounded by $\kappa_2(A|_{\mathcal{R}(A)})$, which, in this case, is the ratio of the largest singular value of A to the smallest positive one. If $x_0 \in \mathcal{R}(A)$ as well, then the solution determined by GMRES is the pseudoinverse solution. If $\mathcal{N}(A) = \mathcal{N}(A^T)$ and the system is not consistent, then, for some b and x_0, the GMRES least-squares problem will necessarily become dangerously ill conditioned before a least-squares solution is reached, despite the theoretical guarantee of no breakdown. However, one may be able to use the condition number of the GMRES least-squares problem to determine when to terminate with nearly the best obtainable accuracy.

If $\mathcal{N}(A) \cap \mathcal{R}(A) = \{0\}$ and the system is consistent, then GMRES will produce a solution without breakdown, even if $\mathcal{N}(A) \neq \mathcal{N}(A^T)$. In this case, the condition number of the GMRES least-squares problem again remains bounded by $\kappa_2(A|_{\mathcal{R}(A)})$, but this may be larger than the ratio of the largest singular value of A to the smallest positive one. Still, this condition number cannot become arbitrarily large through an unfortunate choice of b and x_0.

In some applications in which the system is not consistent, it may be possible to project b onto $\mathcal{R}(A)$. If GMRES can be expected to solve consistent systems reliably, e.g., if $\mathcal{N}(A) = \mathcal{N}(A^T)$ or $\mathcal{N}(A) \cap \mathcal{R}(A) = \{0\}$, then applying GMRES to the consistent system with the projected b will safely yield a least-squares solution of the original inconsistent system.

In practice, the kth GMRES step is obtained by reducing the GMRES least-squares problem to an unconstrained k-dimensional least-squares problem, which is solved through QR factorization. In numerically sound GMRES implementations, singularity or ill conditioning of A is a concern only if it becomes manifested in singularity or ill conditioning of the upper-triangular factors, which may or may not occur before a solution is found. The condition numbers of these factors can be estimated very efficiently using incremental condition estimation (ICE) [4], [5].

Acknowledgment. The authors are grateful to the referees and the editor for their helpful comments and suggestions, which significantly improved the paper.

REFERENCES

[1] E. Anderson, et al., *LAPACK User's Guide*, SIAM, Philadelphia, PA, 1992.
[2] Z. Bai, D. Hu, and L. Reichel, *A Newton Basis GMRES Implementation*, Tech. report 91-03, Department of Mathematics, University of Kentucky, Lexington, KY, April 1991.
[3] V. A. Barker, *Numerical solution of sparse singular systems of equations arising from ergodic Markov chain modeling*, Comm. Statist. Stochastic Models, 5 (1989), pp. 335–381.
[4] C. H. Bischof, *Incremental condition estimation*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 312–322.
[5] C. H. Bischof and P. T. P. Tang, *Robust Incremental Condition Estimation*, Tech. report CS-91-133, LAPACK Working Note 33, Computer Science Department, University of Tennessee, Knoxville, TN, May 1991.
[6] P. N. Brown, *A theoretical comparison of the Arnoldi and GMRES algorithms*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 58–78.
[7] R. W. Freund, *A transpose-free quasi-minimal residual algorithm for non-Hermitian linear systems*, SIAM J. Sci. Comput., 14 (1993), pp. 425–448.
[8] R. W. Freund, G. H. Golub, and N. M. Nachtigal, *Iterative solution of linear systems*, Acta Numerica, 1 (1992), pp. 57–100.
[9] R. W. Freund and M. Hochbruck, *On the use of two QMR algorithms for solving singular systems and applications in Markov chain modeling*, Numer. Linear Algebra Appl., 1 (1994), pp. 403–420.
[10] R. W. Freund and N. M. Nachtigal, *QMR: a quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.

[11] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[12] J. C. Meza, *A Modification to the GMRES Method for Ill-Conditioned Linear Systems*, Tech. report SAND95-8220, Sandia National Laboratories, Livermore, CA, April 1995.
[13] J. C. Meza and W. W. Symes, *Deflated Krylov subspace methods for nearly singular linear systems*, J. Optim. Theory Appl., 72 (1992), pp. 441–457.
[14] R. B. Morgan, *A restarted GMRES method augmented with eigenvectors*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1154–1171.
[15] B. Philippe, Y. Saad, and W. J. Stewart, *Numerical methods in Markov chain modeling*, Oper. Res., 40 (1992), pp. 1156–1179.
[16] Y. Saad and M. H. Schultz, *GMRES: a generalized minimal residual method for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[17] P. N. Swarztrauber and R. A. Sweet, *Efficient fortran subprograms for the solution of elliptic partial differential equations*, ACM Trans. Math. Software, 5 (1979), pp. 352–364.
[18] H. F. Walker, *Implementation of the GMRES method using Householder transformations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 152–163.
[19] H. F. Walker and L. Zhou, *A simpler GMRES*, Numer. Linear Algebra Appl., 1 (1994), pp. 571–581.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 52–65, January 1997

005

STABILITY OF THE DIAGONAL PIVOTING METHOD WITH PARTIAL PIVOTING*

NICHOLAS J. HIGHAM†

Abstract. LAPACK and LINPACK both solve symmetric indefinite linear systems using the diagonal pivoting method with the partial pivoting strategy of Bunch and Kaufman [*Math. Comp.*, 31 (1977), pp. 163–179]. No proof of the stability of this method has appeared in the literature. It is tempting to argue that the diagonal pivoting method is stable for a given pivoting strategy if the growth factor is small. We show that this argument is false in general and give a sufficient condition for stability. This condition is not satisfied by the partial pivoting strategy because the multipliers are unbounded. Nevertheless, using a more specific approach we are able to prove the stability of partial pivoting, thereby filling a gap in the body of theory supporting LAPACK and LINPACK.

Key words. symmetric indefinite matrix, diagonal pivoting method, $\mathrm{LDL}^{\mathrm{T}}$ factorization, partial pivoting, growth factor, numerical stability, rounding error analysis, LAPACK, LINPACK

AMS subject classifications. 65F05, 65G05

PII. S0895479895290371

1. Introduction.

LAPACK is renowned for the numerical reliability of the algorithms it employs. The *LAPACK Users' Guide* [1] states that "almost all the algorithms in LAPACK (as well as LINPACK and EISPACK) are [normwise backward] stable" [1, page 74], and the algorithms not covered by this statement are known to be stable in appropriately weakened senses. The analyses to back up these claims of stability are spread throughout the research literature of the last 35 years. While writing the book *Accuracy and Stability of Numerical Algorithms* [14] we realized that there is no proof in the literature of the stability of the method used in LAPACK and LINPACK for solving symmetric indefinite linear systems. Furthermore, the stability is not a direct consequence of existing results. The purpose of this paper is to prove the stability of the method and thereby to fill a gap in the body of theory supporting LAPACK and LINPACK.

In the remainder of the introduction we briefly describe the method to be analyzed: the diagonal pivoting method with the partial pivoting strategy of Bunch and Kaufman [5].

Let $A \in \mathbb{R}^{n\times n}$ be symmetric. If A 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. Then we can compute the factorization

$$\Pi A \Pi^T = \begin{bmatrix} I_s & 0 \\ CE^{-1} & I_{n-s} \end{bmatrix} \begin{bmatrix} E & 0 \\ 0 & B - CE^{-1}C^T \end{bmatrix} \begin{bmatrix} I_s & E^{-1}C^T \\ 0 & I_{n-s} \end{bmatrix}. \tag{1.1}$$

This process can be repeated recursively on the $(n-s)\times(n-s)$ Schur complement

$$S = B - CE^{-1}C^T.$$

*Received by the editors July 25, 1995; accepted for publication (in revised form) by D. P. O'Leary December 19, 1995. This work was supported by Engineering and Physical Sciences Research Council grants GR/H/52139 and GR/H/94528.

http://www.siam.org/journals/simax/18-1/29037.html

†Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (na.nhigham@na-net.ornl.gov).

The result is a factorization

$$PAP^T = LDL^T, \tag{1.2}$$

where L is unit lower triangular and D is block diagonal with each diagonal block having dimension 1 or 2. This factorization is essentially a symmetric block form of Gaussian elimination, with pivoting, and it costs $n^3/3$ flops[1] (the same cost as Cholesky factorization of a positive definite matrix) plus the cost of determining the permutations Π. This method for computing a block $\mathrm{LDL}^{\mathrm{T}}$ factorization is called the *diagonal pivoting method*. Given the factorization (1.2) of a nonsingular A, a linear system $Ax = b$ is readily solved by substitution and by solving 2×2 linear systems corresponding to any 2×2 diagonal blocks of D.

The strategy for choosing Π is crucial for achieving stability. Bunch and Parlett [7] proposed a complete pivoting strategy, which requires the whole active submatrix to be searched on each stage of the factorization and therefore requires up to $n^3/6$ comparisons. Bunch [3] proved that the diagonal pivoting method with complete pivoting satisfies a backward error bound almost as good as that for Gaussian elimination with complete pivoting. Bunch and Kaufman [5] devised a partial pivoting strategy that searches at most two columns at each stage and so requires only $O(n^2)$ comparisons. The LAPACK driver routines `xSYSV` (simple) and `xSYSVX` (expert) and the LINPACK routines `xSIFA/xSISL` all use the diagonal pivoting method with partial pivoting to solve a linear system with a symmetric (indefinite) coefficient matrix.

To describe the partial pivoting strategy it suffices to define the pivot choice for the first stage of the factorization. Recall that s denotes the size of the pivot block.

ALGORITHM 1 (Bunch–Kaufman partial pivoting strategy). *This algorithm determines the pivot for the first stage of the diagonal pivoting method with partial pivoting applied to a symmetric matrix* $A \in \mathbb{R}^{n\times n}$.

$\alpha := (1 + \sqrt{17})/8 \ (\approx 0.64)$
$\lambda := \|A(2\colon n, 1)\|_\infty$
If $\lambda = 0$ there is nothing to do on this stage of the elimination.
$r := \min\{i \geq 2\colon |a_{i1}| = \lambda\}$
if $|a_{11}| \geq \alpha\lambda$
(1) $s = 1$, $\Pi = I$
else
$\sigma := \left\| \begin{bmatrix} A(1\colon r-1, r) \\ A(r+1\colon n, r) \end{bmatrix} \right\|_\infty$
if $|a_{11}|\sigma \geq \alpha\lambda^2$
(2) $s = 1$, $\Pi = I$
else if $|a_{rr}| \geq \alpha\sigma$
(3) $s = 1$ and choose Π to swap rows and columns 1 and r.
else
(4) $s = 2$ and choose Π to swap rows and columns 2 and r,
so that $|(\Pi A \Pi^T)_{21}| = \lambda$.
end
end

[1] A flop is a floating point addition, subtraction, multiplication, or division.

To understand the partial pivoting strategy it helps to consider the matrix

$$\begin{bmatrix} a_{11} & \dots & \lambda & \dots & \dots & \dots \\ \vdots & & \vdots & & & \\ \lambda & \dots & a_{rr} & \dots & \sigma & \dots \\ \vdots & & \vdots & & & \\ \vdots & & \sigma & & & \\ \vdots & & \vdots & & & \end{bmatrix}$$

and to note that the pivot is one of a_{11}, a_{rr}, and $\left[\begin{smallmatrix} a_{11} & \lambda \\ \lambda & a_{rr} \end{smallmatrix}\right]$ (or, rather, since $\lambda = |a_{r1}|$, this matrix with λ replaced by a_{r1}).

The value of the constant $\alpha = (1+\sqrt{17})/8$ is determined by regarding α as a free parameter and equating a bound for the element growth over two $s = 1$ stages to a bound for the element growth over one $s = 2$ stage; see [5] or [14] for the details.

A growth factor can be defined for the diagonal pivoting method in just the same way as for Gaussian elimination:

$$\rho_n = \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|},$$

where the $a_{ij}^{(k)}$ are the elements of the Schur complements arising in the course of the factorization. From the derivation of the constant α it is easy to show that $\rho_n \le (1 + 1/\alpha)^{n-1} = (2.57)^{n-1}$ for partial pivoting, which is larger than the bound 2^{n-1} for Gaussian elimination with partial pivoting (GEPP). But, it seems that as for GEPP, large element growth is rare in practice [5], [9].

2. Stability of the diagonal pivoting method. Since the growth factor for the diagonal pivoting method with partial pivoting is bounded and is usually small in practice, does it not follow that the method is stable in the same sense as for GEPP? This is a tempting argument, and one that is neither used nor warned against in the existing literature. However, it is easy to show that the argument is false by exhibiting an example where the diagonal pivoting method has a small growth factor but is unstable. An example (not produced by partial pivoting) is, with $n = 3$ and with a 2×2 pivot followed by a 1×1 pivot,

$$\begin{aligned} A &= \begin{bmatrix} 1 & -(1+\epsilon^2) & -\epsilon \\ -(1+\epsilon^2) & 1 & -\epsilon \\ -\epsilon & -\epsilon & -1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & & \\ 0 & 1 & \\ \epsilon^{-1} & \epsilon^{-1} & 1 \end{bmatrix} \begin{bmatrix} 1 & -(1+\epsilon^2) & \\ -(1+\epsilon^2) & 1 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & \epsilon^{-1} \\ & 1 & \epsilon^{-1} \\ & & 1 \end{bmatrix} = LDL^T, \end{aligned} \tag{2.1}$$

where $\epsilon > 0$. The growth factor ρ_n is 1, yet $\|L\|_\infty / \|A\|_\infty$ is unbounded as $\epsilon \to 0$, which suggests that the factorization, however it is computed, may not provide a stable way to solve linear systems $Ax = b$ in finite precision arithmetic. The instability is confirmed by a MATLAB experiment, in which the unit roundoff $u = 2^{-53} \approx 1.1 \times 10^{-16}$. We solved a linear system $Ax = b$, where $b = A\,[1\ 2\ 3]^T$, in two different ways. First, we computed the factorization in (2.1) using the diagonal pivoting method, as specified in (1.1) (with $\Pi = I$), taking a 2×2 pivot on the first step and using GEPP to solve linear systems involving this pivot. For comparison, we evaluated the explicit formulae for the LDL$^{\mathrm{T}}$ factors in (2.1) and used the explicit inverse of $D(1\colon 2, 1\colon 2)$

TABLE 2.1
Backward error for computed solution of indefinite system of order 3.

ϵ	Diagonal pivoting	Explicit factors
10^{-1}	9e−17	6e−16
10^{-2}	5e−17	2e−14
10^{-3}	3e−15	5e−11
10^{-4}	7e−14	4e−9
10^{-5}	6e−13	6e−8
10^{-6}	1e−13	1e−6
10^{-7}	4e−11	1e−7

when solving the linear system involving D. Table 2.1 shows the normwise relative backward error of the computed solution $\widehat{x}$,

$$\begin{aligned}\eta_\infty(\widehat{x}) &:= \min\{\,\epsilon : (A+\Delta A)\widehat{x} = b+\Delta b,\ \ \|\Delta A\|_\infty \le \epsilon\|A\|_\infty,\ \ \|\Delta b\|_\infty \le \epsilon\|b\|_\infty\,\} \\ &= \frac{\|b - A\widehat{x}\|_\infty}{\|A\|_\infty\|\widehat{x}\|_\infty + \|b\|_\infty}\end{aligned}$$

(see [16] or [14, Theorem 7.1] for a proof of the latter equality), which would be of order u for a stable solution method. As ϵ decreases the computations become unstable. We note that stability is obtained if, in (1.1), we take the natural 1×1 pivot a_{11} instead of the ill conditioned 2×2 pivot $A(1\colon 2, 1\colon 2)$; interestingly, though, the 2×2 pivot shares with those chosen by the Bunch–Kaufman partial pivoting strategy the property that it is indefinite. Partial pivoting is stable on this example.

We conclude that a small growth factor is not, by itself, enough to guarantee stability of the diagonal pivoting method. A sufficient condition for stability can be obtained by regarding the block $\mathrm{LDL}^{\mathrm{T}}$ factorization computed by the diagonal pivoting method as a special case of a block LU factorization. Error analysis for block LU factorization is given by Demmel, Higham, and Schreiber [8], and a suitable modification of this analysis gives the following result: if linear systems involving 2×2 pivots are solved in a normwise backward stable fashion then the condition

$$\|L\|_\infty\|D\|_\infty\|L^T\|_\infty \le c_n\|A\|_\infty, \tag{2.2}$$

for a modest constant c_n, is sufficient to ensure that the diagonal pivoting method produces a factorization with a small relative residual and provides computed solutions to linear systems that have a small backward error. Unfortunately, condition (2.2) does not hold for the partial pivoting strategy of Bunch and Kaufman, as is shown by the following example. For $\epsilon > 0$, the diagonal pivoting method with partial pivoting produces the factorization, with $P = I$,

$$A = \begin{bmatrix} 0 & \epsilon & 0 \\ \epsilon & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & & \\ 0 & 1 & \\ 1/\epsilon & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & \epsilon & \\ \epsilon & 0 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1/\epsilon \\ & 1 & 0 \\ & & 1 \end{bmatrix} = LDL^T.$$

As $\epsilon \to 0$, $\|L\|_\infty\|D\|_\infty\|L^T\|_\infty/\|A\|_\infty \to \infty$, and indeed the multipliers are unbounded. Even 1×1 pivots can lead to arbitrarily large elements in L, as the following example with $0 < \epsilon < \alpha$ shows (again, partial pivoting selects $P = I$):

$$A = \begin{bmatrix} \epsilon^2 & \epsilon & \epsilon \\ \epsilon & 0 & 1 \\ \epsilon & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & & \\ 1/\epsilon & 1 & \\ 1/\epsilon & 0 & 1 \end{bmatrix} \begin{bmatrix} \epsilon^2 & & \\ & -1 & \\ & & -1 \end{bmatrix} \begin{bmatrix} 1 & 1/\epsilon & 1/\epsilon \\ & 1 & 0 \\ & & 1 \end{bmatrix} = LDL^T.$$

It is worth emphasizing that large elements in a factor of a matrix do not necessarily imply that the factorization is unstable. For example, in the (point) $\mathrm{LDL}^{\mathrm{T}}$ factorization of a symmetric positive definite matrix A with $D = \mathrm{diag}(d_{ii})$, $d_{ii} > 0$, the ratio $\|L\|_\infty / \|A\|_\infty$ can be arbitrarily large, yet the factorization is guaranteed to be stable. One such example is, with $\epsilon > 0$,

$$A = \begin{bmatrix} \epsilon^2 & \epsilon \\ \epsilon & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \epsilon^{-1} & 1 \end{bmatrix} \begin{bmatrix} \epsilon^2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \epsilon^{-1} \\ 0 & 1 \end{bmatrix}.$$

Our conclusion is that existing results for LU factorization and block LU factorization do not directly imply the stability of the diagonal pivoting method with partial pivoting. Any proof of stability must make use of the particular properties of the partial pivoting strategy.

The only claims of stability that we have found in the literature are in the paper by Bunch, Kaufman, and Parlett [6] and in the *LINPACK Users' Guide* [9, p. 5.19]; in both cases, residual bounds of the form $\|A - \widehat{L}\widehat{D}\widehat{L}^T\|_\infty \le p(n)\rho_n \|A\|_\infty u$ are stated without proof, where p is a polynomial; we prove a result of this form and, in Theorem 4.2, a backward error result for the computed solution of $Ax = b$. We note that much of Bunch's analysis of the diagonal pivoting method in [3] is specific to complete pivoting, so his analysis does not readily yield results for partial pivoting.

In the rest of the paper we present a new analysis to show that partial pivoting is indeed a stable pivoting strategy for the diagonal pivoting method.

3. Background results from error analysis. We collect in this section some standard error analysis results that will be needed later. For our model of floating point arithmetic we take

$$fl(x \mathrm{\ op\ } y) = (x \mathrm{\ op\ } y)(1+\delta), \qquad |\delta| \le u, \quad \mathrm{op} = +, -, *, /, \tag{3.1}$$

where u is the unit roundoff. All the results we quote remain true under a weaker model that accommodates machines without a guard digit [14, section 2.4], provided some of the constants are increased slightly.

We introduce the constant

$$\gamma_n = \frac{nu}{1-nu},$$

which carries with it the implicit assumption that $nu < 1$. Useful properties are (a) $\gamma_m + \gamma_n + \gamma_m\gamma_n \le \gamma_{m+n}$ and (b) if $c \ge 1$ then $c\gamma_n \le \gamma_{cn}$.

Proofs of the following results can be found in [14]. First, for matrix multiplication,

$$fl(AB) = AB + \Delta, \quad |\Delta| \le \gamma_n |A||B|, \quad A \in \mathbb{R}^{m\times n}, \quad B \in \mathbb{R}^{n\times p}.$$

Second, if $T \in \mathbb{R}^{n\times n}$ is a nonsingular triangular matrix and the system $Tx = b$ is solved by substitution then

$$(T + \Delta T)\widehat{x} = b, \qquad |\Delta T| \le \gamma_n |T|. \tag{3.2}$$

Third, if a linear system $Ax = b$, where $A \in \mathbb{R}^{n\times n}$, is solved without breakdown by Gaussian elimination without pivoting, then the computed solution satisfies

$$(A + \Delta A)\widehat{x} = b, \qquad |\Delta A| \le 2\gamma_n |\widehat{L}||\widehat{U}|, \tag{3.3}$$

where $\widehat{L}$ and $\widehat{U}$ are the computed LU factors.

We will use the norm defined by

$$\|A\|_M = \max_{i,j} |a_{ij}|$$

(for which $\|AB\|_M \le n\|A\|_M\|B\|_M$ is the best bound of this form that holds for all $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$).

4. Error analysis.

4.1. 2×2 linear systems. Crucial to the error analysis that follows is a backward error result for the solution of linear systems involving 2×2 pivots. Note that, in the notation of Algorithm 1, the pivot is

$$E = \begin{bmatrix} a_{11} & a_{r1} \\ a_{r1} & a_{rr} \end{bmatrix}, \qquad |a_{r1}| = \lambda.$$

For this subsection and the later analysis, it is convenient to tabulate the conditions that must hold for a 2×2 pivot to be selected:

$$|a_{11}| < \alpha\lambda, \tag{4.1a}$$

$$|a_{11}|\sigma < \alpha\lambda^2, \tag{4.1b}$$

$$|a_{rr}| < \alpha\sigma, \tag{4.1c}$$

$$|a_{11}||a_{rr}| < \alpha^2\lambda^2, \tag{4.1d}$$

where the fourth inequality is a consequence of the previous two (note that (4.1c) implies $\sigma \neq 0$).

Suppose, first, that linear systems $Ex = b$ are solved by GEPP. By (4.1a), $|a_{11}| < \alpha|a_{r1}| < |a_{r1}|$, so GEPP interchanges rows 1 and 2 of E and computes the LU factorization

$$PE = \begin{bmatrix} a_{r1} & a_{rr} \\ a_{11} & a_{r1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \dfrac{a_{11}}{a_{r1}} & 1 \end{bmatrix} \begin{bmatrix} a_{r1} & a_{rr} \\ 0 & a_{r1} - \dfrac{a_{11}a_{rr}}{a_{r1}} \end{bmatrix} = LU.$$

From (3.3), we have the backward error result

$$(PE + \Delta E)\widehat{x} = Pb, \qquad |\Delta E| \le 2\gamma_2 |\widehat{L}||\widehat{U}|.$$

Now

$$|L||U| \le \begin{bmatrix} |a_{r1}| & |a_{rr}| \\ |a_{11}| & \left|\dfrac{a_{11}a_{rr}}{a_{r1}}\right| + \left|a_{r1} - \dfrac{a_{11}a_{rr}}{a_{r1}}\right| \end{bmatrix} \le \begin{bmatrix} |a_{r1}| & |a_{rr}| \\ |a_{11}| & (2\alpha^2+1)|a_{r1}| \end{bmatrix},$$

using (4.1d). It follows that

$$(E + \widetilde{\Delta E})\widehat{x} = b, \qquad |\widetilde{\Delta E}| \le 2\gamma_2 \begin{bmatrix} |a_{11}| & 2|a_{r1}| \\ |a_{r1}| & |a_{rr}| \end{bmatrix} \le 4\gamma_2|E|, \tag{4.2}$$

using the numerical value of α specified in Algorithm 1. Strictly, we should append "$+O(u^2)$" to this bound to account for replacing $|\widehat{L}||\widehat{U}|$ by a bound for $|L||U|$; we omit the second-order term for the moment and reinstate it later. Note that the result (4.2) holds trivially for a 1×1 pivot E.

The main alternative to using GEPP to solve the systems $Ex = b$ is to use the explicit inverse of E, as is done in the implementations of the diagonal pivoting method

with partial pivoting in LAPACK and LINPACK (see the auxiliary routine `xLASYF` in LAPACK and `xSIFA` in LINPACK). In both LAPACK and LINPACK, $Ex = b$ is solved by evaluating

$$x = \frac{1}{a_{r1}\left(\dfrac{a_{11}}{a_{r1}} \cdot \dfrac{a_{rr}}{a_{r1}} - 1\right)} \begin{bmatrix} \dfrac{a_{rr}}{a_{r1}} & -1 \\ -1 & \dfrac{a_{11}}{a_{r1}} \end{bmatrix} b, \tag{4.3}$$

which corresponds to using an explicit formula for the inverse of a 2×2 matrix (or, equivalently, Cramer's rule), with scaling to avoid overflow. The term

$$\mu = \frac{a_{11}}{a_{r1}} \cdot \frac{a_{rr}}{a_{r1}} - 1$$

appears to be a potential source of instability, since for arbitrary a_{11}, a_{r1}, and a_{rr} the relative error in the computed $\widehat{\mu}$ is unbounded. However, by exploiting the condition (4.1d) for a 2×2 pivot, which we rewrite as

$$\frac{|a_{11}||a_{rr}|}{a_{r1}^2} \leq \alpha^2,$$

we can obtain a very satisfactory error bound for $\widehat{\mu}$. Using the model (3.1) we have

$$\widehat{\mu} = \left(\frac{a_{11}}{a_{r1}} \cdot \frac{a_{rr}}{a_{r1}}(1+\delta_1)(1+\delta_2)(1+\delta_3) - 1\right)(1+\delta_4),$$

where $|\delta_i| \leq u$, $i = 1\colon 4$, which implies [14, Lemma 3.1]

$$\widehat{\mu} = \frac{a_{11}}{a_{r1}} \cdot \frac{a_{rr}}{a_{r1}}(1+\theta_4) - (1+\delta_4), \qquad |\theta_4| \leq \gamma_4.$$

Hence

$$\begin{aligned} |\mu - \widehat{\mu}| &\leq \gamma_4 \left(\frac{|a_{11}a_{rr}|}{a_{r1}^2} + 1\right) \leq \gamma_4(\alpha^2 + 1) \\ &\leq \gamma_4 \left(\frac{1+\alpha^2}{1-\alpha^2}\right) |\mu| < 3\gamma_4|\mu|. \end{aligned}$$

It is then straightforward to show that, denoting the matrix in (4.3) by Z,

$$\widehat{x} = (a_{r1}\mu)^{-1}(Z + \Delta Z)b, \qquad |\Delta Z| \leq \gamma_{30}|Z|.$$

Thus $b - E\widehat{x} = -E((a_{r1}\mu)^{-1}\Delta Z)b$, so that

$$\begin{aligned} |b - E\widehat{x}| &\leq \gamma_{30}|E||E^{-1}||b| \\ &\leq \gamma_{30}|E||E^{-1}||E||x| \\ &\leq \gamma_{180}|E||x|, \end{aligned} \tag{4.4}$$

using (A.3). The Oettli–Prager theorem [15], [14, Theorem 7.3] then implies that

$$(E + \Delta E)\widehat{x} = b, \qquad |\Delta E| \leq \gamma_{180}|E|.$$

Again, strictly a second-order term should be added to the bound, this time to account for the fact that $|x|$ rather than $|\widehat{x}|$ appears on the right-hand side of (4.4).

The conclusion is that whether the linear system $Ex = b$ involving the 2×2 pivot is solved by GEPP or by using the explicit inverse, we have

$$(E + \Delta E)\widehat{x} = b, \qquad |\Delta E| \leq \gamma_c|E| \tag{4.5}$$

for an integer constant c. It is worth stressing that such a result does not hold for an arbitrary 2×2 (symmetric) matrix E—we have fully exploited the pivoting conditions in the derivation.

4.2. Componentwise backward error analysis. Now we carry out a componentwise backward error analysis of the diagonal pivoting method. We make only one assumption about the pivoting strategy: that (4.5) holds for the 2×2 pivots. For convenience, we assume, without loss of generality, that no interchanges are needed, which amounts to redefining $A := PAP^T$ in (1.2).

To begin, we consider the first stage of the factorization, using the notation of (1.1). The submatrix $L_{21} = CE^{-1} \in \mathbb{R}^{(n-s)\times s}$ satisfies $L_{21}E = C$ or $EL_{21}^T = C^T$. If l_j is the jth column of L_{21}^T and c_j is the jth column of C^T, then, from (4.5),

$$(E + \Delta E_j)\widehat{l}_j = c_j, \qquad |\Delta E_j| \le \gamma_c |E|.$$

Hence, overall,

$$\widehat{L}_{21}E = C + \Delta C, \qquad |\Delta C| \le \gamma_c |\widehat{L}_{21}||E|. \tag{4.6}$$

We assume that the Schur complement is computed as $S = B - L_{21}C^T$, so that[2]

$$\widehat{S} = B - \widehat{L}_{21}C^T + \Delta S, \qquad |\Delta S| \le \gamma_{s+1}\big(|B| + |\widehat{L}_{21}||C^T|\big). \tag{4.7}$$

The remaining stages of the diagonal pivoting method factorize the Schur complement as $S = L_S D_S L_S^T$, and we assume, inductively, that the computed factors satisfy

$$\widehat{L}_S\widehat{D}_S\widehat{L}_S^T = \widehat{S} + \Delta_S, \qquad |\Delta_S| \le d(n-s,u)\big(|\widehat{S}| + |\widehat{L}_S||\widehat{D}_S||\widehat{L}_S^T|\big),$$

where $d(n-s,u)$ is a constant depending on $n-s$ and u. We therefore have computed factors $\widehat{L}$ and $\widehat{D}$ of A that satisfy

$$\begin{aligned}
\widehat{L}\widehat{D}\widehat{L}^T &:= \begin{bmatrix} I & 0 \\ \widehat{L}_{21} & \widehat{L}_S \end{bmatrix} \begin{bmatrix} E & 0 \\ 0 & \widehat{D}_S \end{bmatrix} \begin{bmatrix} I & \widehat{L}_{21}^T \\ 0 & \widehat{L}_S^T \end{bmatrix} \\
&= \begin{bmatrix} E & E\widehat{L}_{21}^T \\ \widehat{L}_{21}E & \widehat{L}_{21}E\widehat{L}_{21}^T + \widehat{L}_S\widehat{D}_S\widehat{L}_S^T \end{bmatrix} \\
&= \begin{bmatrix} E & (C+\Delta C)^T \\ C + \Delta C & \widehat{L}_{21}E\widehat{L}_{21}^T + \widehat{S} + \Delta_S \end{bmatrix} \\
&= \begin{bmatrix} E & (C+\Delta C)^T \\ C + \Delta C & B + (\widehat{L}_{21}E\widehat{L}_{21}^T - \widehat{L}_{21}C^T) + \Delta S + \Delta_S \end{bmatrix}.
\end{aligned}$$

Now, from (4.6) we have the inequalities

$$|\widehat{L}_{21}E\widehat{L}_{21}^T - \widehat{L}_{21}C^T| \le \gamma_c |\widehat{L}_{21}||E||\widehat{L}_{21}^T|$$

and

$$|\widehat{L}_{21}||C^T| \le (1+\gamma_c)|\widehat{L}_{21}||E||\widehat{L}_{21}^T|. \tag{4.8}$$

Using (4.7) and (4.8) we have

$$|\widehat{S}| \le (1+\gamma_{s+1})(|B| + (1+\gamma_c)|\widehat{L}_{21}||E||\widehat{L}_{21}^T|).$$

[2] If the Schur complement is computed as $S = B - L_{21}EL_{21}^T$ then the same bound (4.9) ensues.

Overall, then, we have

$$\widehat{L}\widehat{D}\widehat{L}^T = A + \Delta A,$$

where $\Delta A_{11} = 0$, $|\Delta A_{21}| \le \gamma_c |\widehat{L}_{21}||E|$, and

$$\begin{aligned}
|\Delta A_{22}| &\le \gamma_c |\widehat{L}_{21}||E||\widehat{L}_{21}^T| + \gamma_{s+1}\bigl(|B| + (1+\gamma_c)|\widehat{L}_{21}||E||\widehat{L}_{21}^T|\bigr) \\
&\quad + d(n-s,u)\bigl((1+\gamma_{s+1})(|B| + (1+\gamma_c)|\widehat{L}_{21}||E||\widehat{L}_{21}^T|) + |\widehat{L}_S||\widehat{D}_S||\widehat{L}_S^T|\bigr) \\
&\le (\gamma_c + d(n-s,u)(1+\gamma_c))|B| + \bigl(\gamma_c(2+\gamma_c) + d(n-s,u)(1+\gamma_c)^2\bigr)|\widehat{L}_{21}||E||\widehat{L}_{21}^T| \\
&\quad + d(n-s,u)|\widehat{L}_S||\widehat{D}_S||\widehat{L}_S^T| \\
&\le (\gamma_c(2+\gamma_c) + d(n-s,u)(1+\gamma_c)^2\bigl)\bigl(|B| + |\widehat{L}_{21}||E||\widehat{L}_{21}^T| + |\widehat{L}_S||\widehat{D}_S||\widehat{L}_S^T|\bigr).
\end{aligned}$$

Hence

$$\widehat{L}\widehat{D}\widehat{L}^T = A + \Delta A, \qquad |\Delta A| \le d(n,u)\bigl(|A| + |\widehat{L}||\widehat{D}||\widehat{L}^T|\bigr), \tag{4.9}$$

where $d(n,u)$ is clearly of the form $p(n)u + O(u^2)$, where p is a linear polynomial.

Now we analyze the substitution stages when the LDL$^{\mathrm{T}}$ factorization is used to solve a linear system $Ax = b$. From (3.2) and (4.5), the computed solutions to the three systems $Ly_1 = b$, $Dy_2 = y_1$, $L^T x = y_2$ satisfy

$$\begin{aligned}
(\widehat{L} + \Delta L_1)\widehat{y}_1 &= b, \qquad |\Delta L_1| \le \gamma_n |\widehat{L}|, \\
(\widehat{D} + \Delta D)\widehat{y}_2 &= \widehat{y}_1, \qquad |\Delta D| \le \gamma_c |\widehat{D}|, \\
(\widehat{L} + \Delta L_2)^T \widehat{x} &= \widehat{y}_2, \qquad |\Delta L_2| \le \gamma_n |\widehat{L}|.
\end{aligned}$$

Thus

$$b = (\widehat{L} + \Delta L_1)(\widehat{D} + \Delta D)(\widehat{L} + \Delta L_2)^T \widehat{x} = (A + \Delta A + \Delta A_2)\widehat{x},$$

where $|\Delta A|$ is bounded in (4.9) and

$$|\Delta A_2| \le \gamma_{2n+c}\, |\widehat{L}||\widehat{D}||\widehat{L}^T| + O(u^2).$$

On bringing back into account the row and column interchanges, we obtain the following result.

THEOREM 4.1. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric and let $\widehat{x}$ be a computed solution to the linear system $Ax = b$ produced by the diagonal pivoting method with any pivoting strategy. If for all linear systems involving 2×2 pivots* (4.5) *holds, then*

$$(A + \Delta A)\widehat{x} = b, \qquad |\Delta A| \le p(n)u\bigl(|A| + P^T|\widehat{L}||\widehat{D}||\widehat{L}^T|P\bigr) + O(u^2), \tag{4.10}$$

where p is a linear polynomial and $PAP^T \approx \widehat{L}\widehat{D}\widehat{L}^T$ is the factorization computed by the diagonal pivoting method.

The bound in (4.10) is analogous to the bound in (3.3) that holds for Gaussian elimination. We have already seen that the assumption (4.5) used in Theorem 4.1 holds for the partial pivoting strategy of Bunch and Kaufman, provided linear systems $Ex = b$ are solved by GEPP or by using the explicit inverse. It is easy to show that this assumption also holds for the complete pivoting strategy of Bunch and Parlett [7] under the same conditions. (Interestingly, for the 2×2 pivots E that arise with the Bunch–Parlett strategy, GEPP applied to $Ex = b$ is identical to Gaussian elimination with complete pivoting.)

4.3. Normwise analysis for partial pivoting. To show that the diagonal pivoting method is stable for a particular pivoting strategy, we need to show that the matrix $|\widehat{L}||\widehat{D}||\widehat{L}^T|$ is suitably bounded. We now focus on partial pivoting. For partial pivoting, $\widehat{L}$ can be arbitrarily large, so stability is not an immediate consequence of Theorem 4.1. We therefore need to look closely at the elements of the matrix $|\widehat{L}||\widehat{D}||\widehat{L}^T|$. For simplicity, we bound the matrix $|L||D||L^T|$ containing the exact factors, which makes only a second-order change to the overall bounds, since $|\widehat{L}||\widehat{D}||\widehat{L}^T| = |L||D||L^T| + O(u)$.

Initially, we examine the contribution from the blocks of L and D produced by the first stage of the factorization. For this more delicate part of the analysis we take full account of the interchanges in our notation. Note that

$$\begin{aligned} |L||D||L^T| &= \begin{bmatrix} I & \\ |L_{21}| & |L_S| \end{bmatrix} \begin{bmatrix} |E| & \\ & |D_S| \end{bmatrix} \begin{bmatrix} I & |L_{21}^T| \\ & |L_S^T| \end{bmatrix} \\ &= \begin{bmatrix} |E| & |E||L_{21}^T| \\ |L_{21}||E| & |L_{21}||E||L_{21}^T| + |L_S||D_S||L_S^T| \end{bmatrix}. \end{aligned} \tag{4.11}$$

We first bound

$$F := |L_{21}||E| = |CE^{-1}||E| \in \mathbb{R}^{(n-s)\times s}.$$

For a 1×1 pivot, F is a vector with elements $|c_i e_{11}^{-1}||e_{11}|$, each of which is trivially bounded by $\max_{i,j} |a_{ij}|$.

Now consider a 2×2 pivot. Algorithm 1 dictates that Π in (1.1) swaps rows and columns 2 and r so that, as noted earlier,

$$E = \begin{bmatrix} a_{11} & a_{r1} \\ a_{r1} & a_{rr} \end{bmatrix}, \qquad |a_{r1}| = \lambda.$$

Using (A.1) and (4.1a), we have

$$\begin{aligned} e_i^T F &\le (e_i^T |C|)|E^{-1}||E| \\ &\le \frac{1}{1-\alpha^2} \begin{bmatrix} \lambda & \sigma \end{bmatrix} \begin{bmatrix} 1+\alpha^2 & \dfrac{2|a_{rr}|}{\lambda} \\ \dfrac{2|a_{11}|}{\lambda} & 1+\alpha^2 \end{bmatrix} \\ &\le \frac{1}{1-\alpha^2} \begin{bmatrix} (1+\alpha^2)\lambda + 2\alpha\sigma & 2|a_{rr}| + (1+\alpha^2)\sigma \end{bmatrix} \\ &\le \frac{\max_{i,j} |a_{ij}|}{1-\alpha^2} \begin{bmatrix} \alpha^2 + 2\alpha + 1 & \alpha^2 + 3 \end{bmatrix} \\ &\le \max_{i,j} |a_{ij}| \begin{bmatrix} 5 & 6 \end{bmatrix}. \end{aligned} \tag{4.12}$$

Next, we need to bound

$$G := |L_{21}||E||L_{21}^T| = |CE^{-1}||E||E^{-1}C^T|.$$

First, consider a 1×1 pivot. In cases (1) and (2) of Algorithm 1 we have

$$g_{ij} = |c_i e_{11}^{-1}||e_{11}||e_{11}^{-1} c_j| = \frac{|a_{i+1,1}||a_{j+1,1}|}{|a_{11}|} \le \frac{\lambda^2}{|a_{11}|} \le \begin{cases} \dfrac{\lambda}{\alpha}, & \text{case (1)}, \\ \dfrac{\sigma}{\alpha}, & \text{case (2)}. \end{cases}$$

In case (3),

$$\begin{aligned} |g_{ij}| &= \frac{|a_{lr}||a_{mr}|}{|a_{rr}|} \qquad (l,m \neq r) \\ &\leq \frac{\sigma^2}{|a_{rr}|} \leq \frac{\sigma}{\alpha}. \end{aligned}$$

For a 1×1 pivot, then, $|g_{ij}| \leq \alpha^{-1} \max_{i,j} |a_{ij}| < 2 \max_{i,j} |a_{ij}|$.

For a 2×2 pivot (case (4) of Algorithm 1), using (A.2) we have

$$\begin{aligned} |g_{ij}| &\leq (e_i^T |C|)\big(|E^{-1}||E||E^{-1}|\big)|C^T|e_j \\ &\leq \frac{3+\alpha^2}{(1-\alpha^2)^2\lambda^2} \begin{bmatrix} \lambda & \sigma \end{bmatrix} \begin{bmatrix} |a_{rr}| & \lambda \\ \lambda & |a_{11}| \end{bmatrix} \begin{bmatrix} \lambda \\ \sigma \end{bmatrix} \\ &= \frac{3+\alpha^2}{(1-\alpha^2)^2\lambda^2} \big(\lambda^2(|a_{rr}|+\sigma) + \sigma(\lambda^2 + |a_{11}|\sigma)\big) \\ &= \frac{3+\alpha^2}{(1-\alpha^2)^2} \left(|a_{rr}| + 2\sigma + \frac{\sigma^2|a_{11}|}{\lambda^2} \right) \\ &\leq \frac{3+\alpha^2}{(1-\alpha^2)^2}(3+\alpha) \max_{i,j} |a_{ij}| \qquad \text{(using (4.1b))} \\ &= 36 \max_{i,j} |a_{ij}|. \end{aligned} \tag{4.13}$$

The remaining blocks of $|L||D||L^T|$ are composed of blocks of L and D that make up $\mathrm{LDL^T}$ factors of Schur complements of A. But every Schur complement satisfies

$$\|S\|_M \leq \rho_n \|A\|_M,$$

where ρ_n is the growth factor. Hence, applying the bounds above recursively to the $(2,2)$ block in (4.11), we deduce the (pessimistic) bound

$$\| \, |L||D||L^T| \, \|_M \leq 36 n \rho_n \|A\|_M. \tag{4.14}$$

We mention in passing that in early drafts of this paper we had a weaker version of (4.5) in which $|E|$ in the bound was replaced by $|E| + |a_{r1}| e_2 e_2^T$. We were still able to obtain a satisfactory bound for $\| \, |L||D||L^T| \, \|_M$, indicating that partial pivoting is somewhat more tolerant of how the 2×2 systems are solved than might be thought from the analysis above.

Using the bound (4.14) in Theorem 4.1 we obtain the following normwise backward stability result for partial pivoting.

THEOREM 4.2. *Let $A \in \mathbb{R}^{n \times n}$ be symmetric and let $\widehat{x}$ be a computed solution to the linear system $Ax = b$ produced by the diagonal pivoting method with the partial pivoting strategy of Bunch and Kaufman, where linear systems involving 2×2 pivots are solved by GEPP or by use of the explicit inverse. Then*

$$(A + \Delta A)\widehat{x} = b, \qquad \|\Delta A\|_M \leq p(n)\rho_n u \|A\|_M + O(u^2), \tag{4.15}$$

where p is a quadratic.

Theorem 4.2 has the same form as Wilkinson's result for GEPP applied to a nonsymmetric system (see, e.g., [14, section 9.2]), though of course the numerical value of ρ_n is usually different for the two methods.

5. Discussion. The backward error matrix ΔA in (4.9) is necessarily symmetric, but that in (4.15) is not, in general. However, we can take ΔA in (4.15) to be symmetric, at the cost of increasing the bound by a factor n, because of the following result of Bunch, Demmel, and Van Loan [4]: if $(A+G)y = b$ then there exists $H = H^T$ such that $(A + H)y = b$ with $\|H\|_2 \leq \|G\|_2$ and $\|H\|_F \leq \sqrt{2}\|G\|_F$.

Sorensen and Van Loan [10, section 5.3.2] modify the Bunch–Kaufman partial pivoting strategy by redefining, in Algorithm 1,

$$\sigma = \|A(:, r)\|_\infty.$$

This small change has the pleasing effect of ensuring that for a positive definite matrix no interchanges are done (and that, as for the Bunch–Kaufman strategy, only 1×1 pivots are used in this case). At the same time it leaves the growth factor bound unchanged, and all our analysis remains valid for this variant.

For sparse symmetric matrices, Duff, Reid, and coauthors compute the block $\mathrm{LDL}^{\mathrm{T}}$ factorization using a pivoting strategy very different from that of Bunch and Kaufman [11], [12], [13]. We describe the strategy in [13] as it applies to the first stage of the factorization: a_{11} is defined to be an acceptable 1×1 pivot, from the point of view of numerical stability, if

$$|a_{11}| \geq \theta \max_{i>1} |a_{i1}|, \tag{5.1}$$

where $\theta \in (0, 1/2]$ is a tolerance; the matrix

$$D_1 = \begin{bmatrix} a_{11} & a_{r1} \\ a_{r1} & a_{rr} \end{bmatrix}$$

is an acceptable 2×2 pivot if

$$\|D_k^{-1}\|_\infty \max\{\, |a_{ij}| : i \neq 1, r;\ j = 1, r \,\} \leq \theta^{-1}. \tag{5.2}$$

From among the acceptable pivots one is chosen that best preserves sparsity, according to some particular sparsity criterion. Conditions (5.1) and (5.2) ensure that $\|L\|_\infty$ is bounded by a multiple of θ^{-1}, which then implies bounds on the growth factor, and hence on $\|D\|_\infty$. The stability of this pivoting strategy is therefore immediate, since (2.2) is satisfied. An interesting contrast is that the Bunch–Kaufman strategy involves a fixed amount of searching for a pivot, and the reasons for its stability are subtle, whereas the Duff et al. strategy more directly forces stability by bounding the multipliers, but gives up the fixed amount of searching of the Bunch–Kaufman strategy.

We emphasize that the aim of this work was to obtain a rigorous backward error bound for the diagonal pivoting method with partial pivoting. The actual performance of the method is affected by the size of the growth factor. More work is needed to investigate the behavior of the growth factor, about which less is known than the growth factor for GEPP. Although the unboundedness of $\|L\|_\infty$ does not preclude backward stability, it does have implications for the practical behavior of the method; see Ashcraft, Grimes, and Lewis [2] for a thorough study for both dense and sparse matrices. Finally, we mention that the implementation of the diagonal pivoting method with partial pivoting in LAPACK 2.0 can be unstable when $\|L\|_\infty$ is large, as pointed out and explained in [2]. The potential instability stems from replacing a symmetric rank-2 update by two rank-1 updates, via the use of an eigendecomposition. This problem will be corrected in a future release of LAPACK.

Appendix. In this appendix we bound three matrix expressions involving a 2×2 pivot from partial pivoting,

$$E = \begin{bmatrix} a_{11} & a_{r1} \\ a_{r1} & a_{rr} \end{bmatrix}, \qquad |a_{r1}| = \lambda.$$

First, we note that

$$|\det(E)| = |a_{r1}^2 - a_{11}a_{rr}| \geq \lambda^2 - \alpha^2\lambda^2 = (1-\alpha^2)\lambda^2,$$

using (4.1d). Hence

$$\begin{aligned} |E^{-1}||E| &\leq \frac{1}{(1-\alpha^2)\lambda^2} \begin{bmatrix} |a_{rr}| & \lambda \\ \lambda & |a_{11}| \end{bmatrix} \begin{bmatrix} |a_{11}| & \lambda \\ \lambda & |a_{rr}| \end{bmatrix} \\ &= \frac{1}{1-\alpha^2} \begin{bmatrix} \dfrac{|a_{11}||a_{rr}|}{\lambda^2} + 1 & \dfrac{2|a_{rr}|}{\lambda} \\ \dfrac{2|a_{11}|}{\lambda} & \dfrac{|a_{11}||a_{rr}|}{\lambda^2} + 1 \end{bmatrix} \\ &\leq \frac{1}{1-\alpha^2} \begin{bmatrix} 1+\alpha^2 & 2\dfrac{|a_{rr}|}{\lambda} \\ 2\dfrac{|a_{11}|}{\lambda} & 1+\alpha^2 \end{bmatrix}, \end{aligned} \tag{A.1}$$

using (4.1d) again. Next,

$$\begin{aligned} |E^{-1}||E||E^{-1}| &\leq \frac{1}{(1-\alpha^2)^2\lambda^2} \begin{bmatrix} 1+\alpha^2 & 2\dfrac{|a_{rr}|}{\lambda} \\ 2\dfrac{|a_{11}|}{\lambda} & 1+\alpha^2 \end{bmatrix} \begin{bmatrix} |a_{rr}| & \lambda \\ \lambda & |a_{11}| \end{bmatrix} \\ &= \frac{1}{(1-\alpha^2)^2\lambda^2} \begin{bmatrix} (3+\alpha^2)|a_{rr}| & (1+\alpha^2)\lambda + 2\dfrac{|a_{11}||a_{rr}|}{\lambda} \\ 2\dfrac{|a_{11}||a_{rr}|}{\lambda} + (1+\alpha^2)\lambda & (3+\alpha^2)|a_{11}| \end{bmatrix} \\ &\leq \frac{1}{(1-\alpha^2)^2\lambda^2} \begin{bmatrix} (3+\alpha^2)|a_{rr}| & (1+3\alpha^2)\lambda \\ (1+3\alpha^2)\lambda & (3+\alpha^2)|a_{11}| \end{bmatrix} \\ &\leq \frac{3+\alpha^2}{(1-\alpha^2)^2\lambda^2} \begin{bmatrix} |a_{rr}| & \lambda \\ \lambda & |a_{11}| \end{bmatrix}. \end{aligned} \tag{A.2}$$

Finally,

$$\begin{aligned} |E||E^{-1}||E| &\leq \frac{1}{1-\alpha^2} \begin{bmatrix} |a_{11}| & \lambda \\ \lambda & |a_{rr}| \end{bmatrix} \begin{bmatrix} 1+\alpha^2 & 2\dfrac{|a_{rr}|}{\lambda} \\ 2\dfrac{|a_{11}|}{\lambda} & 1+\alpha^2 \end{bmatrix} \\ &= \frac{1}{1-\alpha^2} \begin{bmatrix} (3+\alpha^2)|a_{11}| & 2\dfrac{|a_{11}||a_{rr}|}{\lambda} + (1+\alpha^2)\lambda \\ (1+\alpha^2)\lambda + 2\dfrac{|a_{11}||a_{rr}|}{\lambda} & (3+\alpha^2)|a_{rr}| \end{bmatrix} \\ &\leq \frac{1}{1-\alpha^2} \begin{bmatrix} (3+\alpha^2)|a_{11}| & (1+3\alpha^2)\lambda \\ (1+3\alpha^2)\lambda & (3+\alpha^2)|a_{rr}| \end{bmatrix} \\ &\leq \left(\frac{3+\alpha^2}{1-\alpha^2}\right)|E| \leq 6|E|. \end{aligned} \tag{A.3}$$

Acknowledgments. It is a pleasure to thank Philip Gill and Michael Saunders for valuable comments, particularly at early stages of this work. I also thank Jim Bunch, Des Higham, and John Lewis for suggesting improvements to draft manuscripts.

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, second 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*, manuscript, 1995.

[3] J. R. Bunch, *Analysis of the diagonal pivoting method*, SIAM J. Numer. Anal., 8 (1971), pp. 656–680.

[4] J. R. Bunch, J. W. Demmel, and C. F. Van Loan, *The strong stability of algorithms for solving symmetric linear systems*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 494–499.

[5] J. R. Bunch and L. Kaufman, *Some stable methods for calculating inertia and solving symmetric linear systems*, Math. Comp., 31 (1977), pp. 163–179.

[6] J. R. Bunch, L. Kaufman, and B. N. Parlett, *Decomposition of a symmetric matrix*, Numer. Math., 27 (1976), pp. 95–109.

[7] 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.

[8] J. W. Demmel, N. J. Higham, and R. S. Schreiber, *Stability of block LU factorization*, Numer. Linear Algebra Appl., 2 (1995), pp. 173–190.

[9] 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.

[10] J. J. Dongarra, I. S. Duff, D. C. Sorensen, and H. A. van der Vorst, *Solving Linear Systems on Vector and Shared Memory Computers*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1991.

[11] 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.

[12] I. S. Duff and J. K. Reid, *MA*27*—A set of Fortran Subroutines for Solving Sparse Symmetric Sets of Linear Equations*, Tech. Report AERE R10533, AERE Harwell Laboratory, Her Majesty's Stationary Office, London, July 1982.

[13] 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.

[14] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1996.

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

[16] J. L. Rigal and J. Gaches, *On the compatibility of a given solution with the data of a linear system*, J. Assoc. Comput. Mach., 14 (1967), pp. 543–548.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 66–73, January 1997

006

ESTIMATING THE SUPPORT OF A SCALING VECTOR*

WASIN SO† AND JIANZHONG WANG†

Abstract. An estimate is given for the support of each component function of a compactly supported scaling vector satisfying a matrix refinement equation with finite number of terms. The estimate is based on the highest and lowest degrees of each polynomial in the corresponding matrix symbol. Only basic techniques from matrix theory are involved in the derivation.

Key words. support, scaling vector, multiwavelets

AMS subject classifications. 15A39, 42C15

PII. S0895479895281460

1. Introduction. In this paper we are interested in measurable functions from the real $\mathbf{R}$ to the complex $\mathbf{C}$; two functions are equal if they are identical almost everywhere. Let r be a positive integer and $F = [f_1 \ \dots \ f_r]^T$ be a complex vector-valued function on $\mathbf{R}$, where ^{T} denotes the transpose of a matrix. A point $t \in \mathbf{R}$ is called a *support point* of F if the measure of the intersection $\{x : F(x) \neq 0\} \cap (t-\epsilon, t+\epsilon)$ is not zero for any $\epsilon > 0$. The support of F, denoted by $\mathrm{supp}(F)$, is defined as the convex hull of the set of support points of F. Hence equal functions have the same supports, and the support of a nonzero function is always a close interval with positive length. Note that, in the literature of wavelet theory with $r = 1$, the support of a scaling function is always taken to be a closed interval because of the result in [5] (also see [1, p. 252]).

Recent interest in multiwavelets led to the study of scaling vector $\Phi = [\phi_1 \ \dots \ \phi_r]^T$, which is a vector-valued function satisfying a matrix refinement equation (MRE) with a finite number of terms

$$\Phi(x) = \sum_{k=0}^{N} C_k \Phi(2x - k), \tag{1}$$

where C_k's are $r \times r$ matrices. In applications, shortly supported multiwavelets are always desired. Support of multiwavelets can be obtained easily from the support of the corresponding scaling vectors. Hence it is useful to estimate the support of scaling vectors from the defining MRE. However, the determination of the support of a scaling vector is not straightforward. In [3], Heil and Colella observed that $\mathrm{supp}(\Phi) \subset [0, N]$ if Φ is compactly supported. But this estimate is too crude, as the following example, due to Geronimo, Hardin, and Massopust [2], shows.

Example. Let $\Phi = [\phi_1 \ \phi_2]^T$ be a scaling vector satisfying the MRE (1) with matrix coefficients

$$C_0 = \frac{1}{20}\begin{bmatrix} 12 & 16\sqrt{2} \\ -\sqrt{2} & -6 \end{bmatrix}, \quad C_1 = \frac{1}{20}\begin{bmatrix} 12 & 0 \\ 9\sqrt{2} & 20 \end{bmatrix},$$

* Received by the editors February 9, 1995; accepted for publication (in revised form) by K. Sigmon December 20, 1995.
http://www.siam.org/journals/simax/18-1/28146.html
† Department of Mathematical and Information Sciences, Sam Houston State University, Huntsville, TX 77341 (mth_wso@shsu.edu, mth_jxw@shsu.edu). Both authors were partially supported by National Science Foundation grant DMS-9503282.

$$C_2 = \frac{1}{20}\begin{bmatrix} 0 & 0 \\ 9\sqrt{2} & -6 \end{bmatrix}, \quad C_3 = \frac{1}{20}\begin{bmatrix} 0 & 0 \\ -\sqrt{2} & 0 \end{bmatrix}.$$

Note that $\text{supp}(\phi_1) = [0,1]$, $\text{supp}(\phi_2) = [0,2]$, and so $\text{supp}(\Phi) = [0,2] \neq [0,3]$.

An explanation is the existence of nilpotent matrices. Note that C_3 in the above example is nilpotent. In [6], Massopust, Ruch, and Van Fleet showed that $\text{supp}(\Phi) \subset [0, N - \frac{1}{2^r-1}]$ if C_N is nilpotent, and $\text{supp}(\Phi) \subset [\frac{1}{2^r-1}, N]$ if C_0 is nilpotent. However, such improved estimates are still not good enough to explain the above example.

In this paper, we give an estimate for each componentwise support $\text{supp}(\phi_i)$ and hence the global support $\text{supp}(\Phi)$. Sufficient conditions are given for these estimates to be achieved. The rest of the paper is organized as follows. Our main results are stated in section 2 with an illustration. Proofs are given in section 3. Section 4 is devoted to the study of the global support of a scaling vector.

2. Componentwise support of a scaling vector. For the rest of the paper, let $\Phi = [\phi_1 \ \dots \ \phi_r]^T$ be a compactly supported scaling vector satisfying the MRE (1). In this section we are interested in estimating the support $\text{supp}(\phi_i)$ for $1 \leq i \leq r$. To this end, we define the associated matrix symbol by

$$P(z) = \sum_{k=0}^{N} C_k z^k,$$

which is an $r \times r$ matrix with polynomial entries. Let $h(i,j)$ (resp., $l(i,j)$) be the highest (resp., lowest) degree of the (i,j) entry of $P(z)$. We adopt the convention that the highest (resp., lowest) degree of the zero polynomial is $-\infty$ (resp., ∞).

I_k denotes the $k \times k$ identity matrix and e_k denotes the kth column of the identity matrix whose dimension is determined from the context. For positive integers a, b, E_{ab} denotes the matrix $e_a e_b^T$.

Let $\mathcal{J}$ be the set of all integer sequences $J = (j_1, \dots, j_r)$, where $1 \leq j_1, \dots, j_r \leq r$. For each $J = (j_1, \dots, j_r) \in \mathcal{J}$, define

$$E_J = 2I_r - E_{1j_1} - \cdots - E_{rj_r},$$

$$h_J = [h(1,j_1) \ \cdots \ h(r,j_r)]^T, \qquad \text{and} \qquad l_J = [l(1,j_1) \ \cdots \ l(r,j_r)]^T.$$

Note that E_J is always invertible (see Lemma 3.2).

THEOREM 2.1. *For $1 \leq i \leq r$, the support of ϕ_i is a finite closed interval $[L_i, R_i]$, where*

$$R_i \leq \max\left\{e_i^T E_J^{-1} h_J : J \in \mathcal{J}\right\}$$

and

$$L_i \geq \min\left\{e_i^T E_J^{-1} l_J : J \in \mathcal{J}\right\}.$$

In Theorem 2.1, both maximization and minimization are with respect to the set $\mathcal{J}$ which has r^r elements. In order to reduce the complexity we introduce the following concepts. For each $J = (j_1, \dots, j_r) \in \mathcal{J}$ and $1 \leq i \leq r$, define a new integer sequence $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_t)$ satisfying the following conditions:

1. $1 \leq t \leq r$,
2. $\gamma_0 = i$,

3. $\gamma_k = j_{\gamma_{(k-1)}}$ for $k = 1, \ldots, t$,
4. $\gamma_0, \ldots, \gamma_{t-1}$ are distinct,
5. $\gamma_t = \gamma_{s-1}$ for some $1 \le s \le t$.

The existence of γ, s, and t is clear, and they are uniquely determined by the sequence $J = (j_1, \ldots, j_r)$ and the integer i. As an example, take $r = 4$. If $J = (3, 2, 4, 3)$ and $i = 1$, then $\gamma = (1, 3, 4, 3)$, $t = 3$, and $s = 2$. If $J = (3, 2, 4, 3)$ and $i = 2$, then $\gamma = (2, 2)$, $t = 1$, and $s = 1$.

For fixed i, let Γ_i be the collection of all such γ's. Let s and t be the numbers corresponding to a given $\gamma \in \Gamma_i$. Define a $t \times t$ matrix by

$$A_\gamma = 2I_t - E_{12} - E_{23} - \cdots - E_{(t-1)t} - E_{ts}.$$

Note that $A_\gamma = E_J$ for $J = (2, 3, \ldots, t-1, s)$ and so A_γ is invertible (see Lemma 3.2). Define

$$h_\gamma = [h(\gamma_0, \gamma_1)\ h(\gamma_1, \gamma_2)\ \ldots\ h(\gamma_{t-1}, \gamma_t)]^T$$

and

$$l_\gamma = [l(\gamma_0, \gamma_1)\ l(\gamma_1, \gamma_2)\ \ldots\ l(\gamma_{t-1}, \gamma_t)]^T.$$

THEOREM 2.2. *For $1 \le i \le r$, the support of ϕ_i is a finite closed interval $[L_i, R_i]$, where*

$$R_i \le \max\left\{e_1^T A_\gamma^{-1} h_\gamma : \gamma \in \Gamma_i\right\}$$

and

$$L_i \ge \min\left\{e_1^T A_\gamma^{-1} l_\gamma : \gamma \in \Gamma_i\right\}.$$

In Theorem 2.2, both maximization and minimization are with respect to the set Γ_i. The number of elements in Γ_i is $\sum_{k=0}^{r-1} \binom{r-1}{k}(k+1)!$ which can be proved to be equal to the integral part of the positive number $(r-1)!(r-1)e + 1$, where e is the base of natural logarithm. Hence the complexity of the optimization is reduced to $(r-1)!(r-1)e + 1$ from r^r in Theorem 2.1.

Using the classical adjoint formula for a matrix inverse [4, p. 20], it is not hard to see that the first row of A_γ^{-1} is

$$e_1^T A_\gamma^{-1} = \left[\frac{1}{2} \quad \cdots \quad \frac{1}{2^{s-1}} \quad \left(\frac{2^t}{2^t - 2^{s-1}}\right)\frac{1}{2^s} \quad \cdots \quad \left(\frac{2^t}{2^t - 2^{s-1}}\right)\frac{1}{2^t}\right].$$

Therefore Theorem 2.2 can be restated explicitly as follows.

THEOREM 2.3. *For $1 \le i \le r$, the support of ϕ_i is a finite closed interval $[L_i, R_i]$, where*

$$R_i \le \max_{\gamma \in \Gamma_i} \left\{\sum_{k=1}^{s-1} \frac{1}{2^k} h(\gamma_{(k-1)}, \gamma_k) + \frac{2^t}{2^t - 2^{s-1}} \sum_{k=s}^{t} \frac{1}{2^k} h(\gamma_{(k-1)}, \gamma_k)\right\}$$

and

$$L_i \ge \min_{\gamma \in \Gamma_i} \left\{\sum_{k=1}^{s-1} \frac{1}{2^k} l(\gamma_{(k-1)}, \gamma_k) + \frac{2^t}{2^t - 2^{s-1}} \sum_{k=s}^{t} \frac{1}{2^k} l(\gamma_{(k-1)}, \gamma_k)\right\}.$$

A family $\{f_i\}$ of functions on $\mathbf{R}$ is *locally linearly independent* if $\sum_i c_i f_i(x) = 0$ on any nontrivial interval (a, b) and implies $c_i = 0$ for all i for which $\operatorname{supp}(f_i) \cap (a, b) \neq \emptyset$. $\Phi = [\phi_1, \ldots, \phi_r]^T$ is called a *locally linearly independent* scaling vector if the family $\{\phi_j(x - k) : 1 \leq j \leq r, k \in \mathbf{Z}\}$ is locally linearly independent. In this case, the family $\{\phi_j(2x - k) : 1 \leq j \leq r, k \in \mathbf{Z}\}$ is also locally linearly independent. This fact will be used in Lemma 3.4.

THEOREM 2.4. *If Φ is a locally linearly independent scaling vector then all inequalities become equalities in Theorems* 2.1*,* 2.2*, and* 2.3*.*

Choosing $r = 2$ in Theorem 2.3 yields

$$R_1 \leq \max\left\{h(1,1),\ \ \frac{2}{3}h(1,2) + \frac{1}{3}h(2,1),\ \ \frac{1}{2}h(1,2) + \frac{1}{2}h(2,2)\right\},$$

$$R_2 \leq \max\left\{h(2,2),\ \ \frac{2}{3}h(2,1) + \frac{1}{3}h(1,2),\ \ \frac{1}{2}h(2,1) + \frac{1}{2}h(1,1)\right\},$$

$$L_1 \geq \min\left\{l(1,1),\ \ \frac{2}{3}l(1,2) + \frac{1}{3}l(2,1),\ \ \frac{1}{2}l(1,2) + \frac{1}{2}l(2,2)\right\},$$

and

$$L_2 \geq \min\left\{l(2,2),\ \ \frac{2}{3}l(2,1) + \frac{1}{3}l(1,2),\ \ \frac{1}{2}l(2,1) + \frac{1}{2}l(1,1)\right\}.$$

As an illustration, we use these formulas to estimate the support of the scaling vector mentioned in the example of section 1. The highest and lowest degree matrices are, respectively, $h = \left[\begin{smallmatrix} 1 & 0 \\ 3 & 2 \end{smallmatrix}\right]$ and $l = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right]$. Hence $0 \leq L_1 \leq R_1 \leq 1$ and $0 \leq L_2 \leq R_2 \leq 2$. Furthermore, Φ is known to be locally linearly independent [2] and so we have $\operatorname{supp}(\phi_1) = [L_1, R_1] = [0, 1]$ and $\operatorname{supp}(\phi_2) = [L_2, R_2] = [0, 2]$.

3. Proofs. We need two lemmas for the proof of Theorem 2.1.

LEMMA 3.1. *Let $\{f_i\}$ be a family of functions on $\mathbf{R}$. Then*

$$\operatorname{supp}\left(\sum_i c_i f_i\right) \subset \mathit{conv}\left(\cup_i \operatorname{supp}(f_i)\right),$$

where "conv" denotes the convex hull of a set.

LEMMA 3.2. *For $J \in \mathcal{J}$, the matrix E_J is invertible and its inverse has nonnegative entries.*

Proof. Let $E = E_{1j_r} + \cdots + E_{rj_r}$. Note that $\|E\| = 1$ where $\|\cdot\|$ is the maximum row sum norm. Then $E_J = 2I_r - E$ is invertible and actually

$$E_J^{-1} = \sum_{k=0}^{\infty} \frac{1}{2^{k+1}} E^k,$$

which has nonnegative entries because E has nonnegative entries. ∎

We are ready to prove Theorem 2.1.

Proof of Theorem 2.1. For each $1 \le i \le r$, using the MRE (1), we have

$$\begin{aligned}\phi_i(x) &= \sum_{k=0}^{N}\sum_{j=1}^{r} C_k(i,j)\phi_j(2x-k)\\ &= \sum_{j=1}^{r}\sum_{k=0}^{N} C_k(i,j)\phi_j(2x-k)\\ &= \sum_{j=1}^{r}\sum_{k=l(i,j)}^{h(i,j)} C_k(i,j)\phi_j(2x-k),\end{aligned}$$

where $C_k(i,j)$ is the (i,j) entry of the matrix C_k. Since ϕ_i has compact support, we let $\operatorname{supp}(\phi_i) = [L_i, R_i]$. By Lemma 3.1, we have

$$\begin{aligned}[L_i, R_i] &\subset \operatorname{conv}\left(\cup_{j=1}^{r} \operatorname{supp}\left(\sum_{k=l(i,j)}^{h(i,j)} C_k(i,j)\phi_j(2x-k)\right)\right)\\ &= \operatorname{conv}\left(\cup_{j=1}^{r}\left[\frac{1}{2}(L_j + l(i,j)), \frac{1}{2}(R_j + h(i,j))\right]\right).\end{aligned}$$

Hence we have

$$2R_i \le \max\{R_j + h(i,j) : 1 \le j \le r\}$$

and

$$2L_i \ge \min\{L_j + l(i,j) : 1 \le j \le r\}.$$

For each $1 \le i \le r$, there exist integers $1 \le j_1, \ldots, j_r \le r$ such that

$$2R_i \le R_{j_i} + h(i, j_i).$$

In matrix form,

$$E_J \begin{bmatrix} R_1 \\ \vdots \\ R_r \end{bmatrix} = \left(2I - \sum_{t=1}^{r} E_{tj_t}\right)\begin{bmatrix} R_1 \\ \vdots \\ R_r \end{bmatrix} \le \begin{bmatrix} h(1,j_1) \\ \vdots \\ h(r,j_r) \end{bmatrix} = h_J,$$

where $J = (j_1, \ldots, j_r)$. By Lemma 3.2, E_J^{-1} is a nonnegative matrix and so

$$\begin{bmatrix} R_1 \\ \vdots \\ R_r \end{bmatrix} \le E_J^{-1} h_J.$$

Hence

$$R_i \le e_i^T E_J^{-1} h_J \le \max_{J \in \mathcal{J}} e_i^T E_J^{-1} h_J.$$

Similarly, the lower bound for L_i is obtained. □

LEMMA 3.3. *Let p be a permutation on $\{1,\dots,r\}$ and P be the $r\times r$ matrix associated with p. Then $P^{-1}=P^T, e_k^TP=e_{p^{-1}(k)}^T, P^TE_{ab}P=E_{p^{-1}(a)p^{-1}(b)}$, and $[v_1\ v_2\ \dots\ v_r]P=[v_{p(1)}\ v_{p(2)}\ \dots\ v_{p(r)}]$.*

Finally we give the proof of Theorem 2.2.

Proof of Theorem 2.2. Given $J\in\mathcal{J}$ and $1\le i\le r$, let γ,s,t be the corresponding sequence and numbers defined in section 2. It suffices to prove that

$$e_i^TE_J^{-1}h_J=e_1^TA_\gamma^{-1}h_\gamma.$$

Take a permutation p on $\{1,\dots,n\}$ such that $p(k)=\gamma_{(k-1)}$ for $k=1,\dots,t$. Such a permutation exists because the integers $\gamma_0,\dots,\gamma_{t-1}$ are distinct. Using Lemma 3.3, we have $P^Te_{\gamma_{(k-1)}}=e_{p^{-1}(\gamma_{(k-1)})}=e_k$ for $k=1,\dots,t$. It follows that $e_i^TP=e_{p^{-1}(i)}^T=e_1^T$ because $p(1)=\gamma_0=i$, $P^Th_J=\left[\begin{smallmatrix}h_\gamma\\ *\end{smallmatrix}\right]$, and

$$\begin{aligned}
P^TE_JP&=P^T\left(2I_r-\sum_{k=1}^r E_{kj_k}\right)P\\
&=2I_r-P^T\left(\sum_{k=1}^t E_{\gamma_{(k-1)}\gamma_k}+\sum_{k\notin\gamma}E_{kj_k}\right)P\\
&=2I_r-\sum_{k=1}^t E_{p^{-1}(\gamma_{k-1})p^{-1}(\gamma_k)}-\sum_{k\notin\gamma}E_{p^{-1}(k)p^{-1}(j_k)}\\
&=2I_r-\sum_{k=1}^t E_{k\ (k+1)}-\sum_{k>t}E_{kj_k}\\
&=\begin{bmatrix}A_\gamma & 0\\ * & *\end{bmatrix}.
\end{aligned}$$

Finally,

$$\begin{aligned}
e_i^TE_J^{-1}h_J&=(e_i^TP)\,(P^TE_J^{-1}P)\,(P^Th_J)\\
&=(e_i^TP)\,(P^TE_JP)^{-1}\,(P^Th_J)\\
&=e_1^T\begin{bmatrix}A_\gamma & 0\\ * & *\end{bmatrix}^{-1}\begin{bmatrix}h_\gamma\\ *\end{bmatrix}\\
&=e_1^TA_\gamma^{-1}h_\gamma.\qquad\square
\end{aligned}$$

LEMMA 3.4. *Let $\{f_1,\dots,f_n\}$ be a family of locally linearly independent functions on $\mathbf{R}$ such that $\operatorname{supp}(f_i)=[a_i,b_i]$, where $a_i<b_i$. Then*

$$\operatorname{supp}\left(\sum_{i=1}^n c_if_i\right)=[a,b]\,,$$

where $a=\min\{a_i:c_i\neq0\}$ and $b=\max\{b_i:c_i\neq0\}$.

Proof. Let $a_l=\min\{a_i:c_i\neq0\}$ and $b_h=\max\{b_i:c_i\neq0\}$. By Lemma 3.1, $\sum_{i=1}^n c_if_i$ is compactly supported and

$$\operatorname{supp}\left(\sum_{i=1}^n c_if_i\right)=[a,b]\subset[a_l,b_h]=\operatorname{conv}\left(\cup_i\operatorname{supp}(f_i)\right).$$

It remains to show that $a = a_l$ and $b = b_h$. Assume the contrary, that $b < b_h$. Then $\sum_{i=1}^{n} c_i f_i(x) = 0$ on $(b_h - \epsilon, b_h)$ for $0 < \epsilon < \min_i \left\{ \frac{b_i - a_i}{2} \right\}$. Note that $[a_h, b_h] \cap (b_h - \epsilon, b_h) \neq \emptyset$. By the local linear independence of $\{f_i\}$, $c_h = 0$, which is impossible by the definition of b_h. The argument for $a = a_l$ is similar. □

Proof of Theorem 2.4. It suffices to give the proof involving Theorem 2.1. For each $1 \leq i \leq r$, using the MRE (1), we have

$$\begin{aligned}\phi_i(x) &= \sum_{k=0}^{N} \sum_{j=1}^{r} C_k(i,j)\phi_j(2x-k) \\ &= \sum_{j=1}^{r} \sum_{k=0}^{N} C_k(i,j)\phi_j(2x-k) \\ &= \sum_{j=1}^{r} \sum_{k=l(i,j)}^{h(i,j)} C_k(i,j)\phi_j(2x-k),\end{aligned}$$

where $C_k(i,j)$ is the (i,j) entry of the matrix C_k. Since ϕ_i has compact support, we let $\operatorname{supp}(\phi_i) = [L_i, R_i]$. By Lemma 3.4, we have

$$\begin{aligned}[L_i, R_i] &= \operatorname{conv}\left(\cup_{j=1}^{r} \operatorname{supp}\left(\sum_{k=l(i,j)}^{h(i,j)} C_k(i,j)\phi_j(2x-k) \right) \right) \\ &= \operatorname{conv}\left(\cup_{j=1}^{r} \left[\frac{1}{2}(L_j + l(i,j)), \frac{1}{2}(R_j + h(i,j)) \right] \right).\end{aligned}$$

The rest of the proof is exactly the same as the proof of Theorem 2.1 with the modification that all inequalities are changed to equalities. □

4. Global support of a scaling vector. In this section we are interested in the global support $\operatorname{supp}(\Phi)$ of Φ satisfying the MRE (1). From the last section we know that $\operatorname{supp}(\phi_i) = [L_i, R_i]$ for $1 \leq i \leq r$. Hence $\operatorname{supp}(\Phi) = [L, R]$, where $R = \max\{R_i : 1 \leq i \leq r\}$ and $L = \min\{L_i : 1 \leq i \leq r\}$. Theorem 2.3 gives the estimates as

$$R \leq \max_{1 \leq i \leq r} \max_{\gamma \in \Gamma_i} \left\{ \sum_{k=1}^{s-1} \frac{1}{2^k} h(\gamma_{k-1}, \gamma_k) + \frac{2^t}{2^t - 2^{s-1}} \sum_{k=s}^{t} \frac{1}{2^k} h(\gamma_{k-1}, \gamma_k) \right\}$$

and

$$L \geq \min_{1 \leq i \leq r} \min_{\gamma \in \Gamma_i} \left\{ \sum_{k=1}^{s-1} \frac{1}{2^k} l(\gamma_{k-1}, \gamma_k) + \frac{2^t}{2^t - 2^{s-1}} \sum_{k=s}^{t} \frac{1}{2^k} l(\gamma_{k-1}, \gamma_k) \right\}.$$

THEOREM 4.1. (i) *If C_N is a nilpotent matrix of index m, i.e., $(C_N)^m = 0$, then $R \leq N - \frac{1}{2^m - 1}$.*

(ii) *If C_0 is a nilpotent matrix of index m, i.e., $(C_0)^m = 0$, then $L \geq \frac{1}{2^m - 1}$.*

Proof. Using Lemma 3.1, it is not hard to see that $\operatorname{supp}(\Phi) = \operatorname{supp}(A\Phi)$ for any invertible matrix A.

(i) Without loss of generality, we can assume that C_N is reduced to the Jordan form $J_m(0) \oplus \cdots$ where $J_m(0)$ is a lower triangular Jordan block with largest size. Hence the highest degree matrix satisfies

$$h \leq (N-1)One(r) + C_N,$$

where $One(r)$ is an $r \times r$ matrix with all entries equal to 1. Now it is not hard to see that the maximum is attained at $i = m$ and $\gamma = (m, m-1, \ldots, 1, m)$. Actually, the maximum is equal to

$$\frac{2^m}{2^m - 1} \sum_{k=1}^{m} \frac{1}{2^k} h(\gamma_{k-1}, \gamma_k) \leq N - \frac{1}{2^m - 1}.$$

(ii) Without loss of generality, we can assume that C_0 is reduced to the Jordan form $J_m(0) \oplus \cdots$ where $J_m(0)$ is a lower triangular Jordan block with largest size. Hence the lowest degree matrix satisfies

$$l \geq One(r) - C_0.$$

Now it is not hard to see that the minimum is attained at $i = m$ and $\gamma = (m, m-1, \ldots, 1, m)$. Actually, the minimum is equal to

$$\frac{2^m}{2^m - 1} \sum_{k=1}^{m} \frac{1}{2^k} l(\gamma_{k-1}, \gamma_k) \geq \frac{1}{2^m - 1}. \qquad \square$$

Setting $m = r$, we obtain the result of Massopust, Ruch, and Van Fleet mentioned in the introduction.

COROLLARY 4.2. (i) *If C_N is nilpotent, then* $\operatorname{supp}(\Phi) \subset [0, N - \frac{1}{2^r - 1}]$.

(ii) *If C_0 is nilpotent, then* $\operatorname{supp}(\Phi) \subset [\frac{1}{2^r - 1}, N]$.

REFERENCES

[1] I. DAUBECHIES, *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA, 1992.

[2] J. S. GERONIMO, D. P. HARDIN, AND P. R. MASSOPUST, *Fractal functions and wavelet expansions based on several scaling functions*, J. Approx. Theory, 78 (1994), pp. 373–401.

[3] C. HEIL AND D. COLELLA, *Matrix refinement equations: Existence and uniqueness*, J. Fourier Anal. Appl., 2 (1996), pp. 363–377.

[4] R. HORN AND C. JOHNSON, *Matrix Analysis*, Cambridge University Press, London, 1989.

[5] P. G. LEMARIE AND G. MALGOUYRES, *Support des fonctions de base daus une analyse multiresolution*, C. R. Acad. Sci. Paris Ser. I Math., 313 (1991), pp. 377–380.

[6] P. MASSOPUST, D. RUCH, AND P. VAN FLEET, *On the support properties of scaling vectors*, Appl. Comp. Harmonic Anal., 3 (1996), pp. 229–238.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 74–82, January 1997

007

DETERMINANT OF THE SUM OF A SYMMETRIC AND A SKEW-SYMMETRIC MATRIX*

NATÁLIA BEBIANO†, CHI-KWONG LI‡, AND JOÃO DA PROVIDÊNCIA§

In memory of Professor Robert Thompson.

Abstract. The set of all possible determinant values of the sum of a (complex) symmetric matrix and a skew-symmetric matrix with prescribed singular values is determined. This set can also be viewed as the best containment region for the determinant of a square matrix X in terms of the singular values of its symmetric and skew-symmetric parts. The technique is extended to study the principal minors of X. Similar problems for real matrices are considered.

Key words. determinant, (skew-)symmetric matrix, majorization

AMS subject classifications. 15A15, 15A45

PII. S0895479895293429

1. Introduction. The determinant of matrices is a very useful tool in both pure and applied mathematics, and its study has a long history (e.g., see [C] and [Mu]). Due to motivations arising from theory and applications, there has been a great deal of interest in estimating $\det(A+B)$ based on information of the square matrices A and B and finding bounds for $\det(X)$ based on partial information of the square matrix X; e.g., see [HJ, section 7.8], [F], [LM], [Ma], [Mi], and [O]. The purpose of this paper is to determine the set of all possible determinant values of the sum of a (complex) symmetric matrix and a skew-symmetric matrix with prescribed singular values. This set can also be viewed as the best containment region for $\det(X)$ in terms of the singular values of $(X+X^t)/2$ and $(X-X^t)/2$. (See [HJ, Chapter 3] for the definition and properties of singular values.)

A complete answer to our question for complex matrices is given in section 2. Our technique is extended to estimate the principal minors of matrices. Similar problems for real matrices are studied in section 3. Some remarks and open problems are mentioned in section 4. In our discussion, we shall use the following notation:

$\mathbb{F}$: the complex field $\mathbb{C}$ or the real field $\mathbb{R}$,
$M_n(\mathbb{F})$: algebra of $n \times n$ matrices over $\mathbb{F}$,
$\{E_{11}, E_{12}, \ldots, E_{nn}\}$: the standard basis of $M_n(\mathbb{F})$,
$S_n(\mathbb{F})$: the set of $n \times n$ symmetric matrices over $\mathbb{F}$,
$K_n(\mathbb{F})$: the set of $n \times n$ skew-symmetric matrices over $\mathbb{F}$,
$\mathcal{U}_n$: the group of $n \times n$ (complex) unitary matrices,
$\mathcal{O}_n$: the group of $n \times n$ (real) orthogonal matrices,
$\mathrm{diag}(d_1, \ldots, d_n)$: the diagonal matrix with $d_1, \ldots, d_n$ as the diagonal entries.

We shall always assume $n \geq 2$ to avoid trivial consideration.

* Received by the editors October 24, 1995; accepted for publication by T. Ando December 21, 1995. This research was supported by a NATO grant.
http://www.siam.org/journals/simax/18-1/29342.html
† Departamento de Matemática, Universidade de Coimbra, 3000 Coimbra, Portugal (bebiano@ciuc2.uc.pt).
‡ Department of Mathematics, The College of William and Mary, Williamsburg, VA 23187 (ckli@math.wm.edu).
§ Departamento de Física, Universidade de Coimbra, 3000 Coimbra, Portugal (fcfiproviden@ciuc2.uc.pt).

2. Results on complex matrices. The following facts (e.g., see [HJ, Chapter 3] and [Y]) are important to our discussion:

(2.1) every complex symmetric matrix with singular values $\alpha_1 \geq \cdots \geq \alpha_n$ can be written as $U^t(\sum_{i=1}^n \alpha_i E_{ii})U$ for some $U \in \mathcal{U}_n$;

(2.2) the singular values $\beta_1 \geq \cdots \geq \beta_n$ of a complex skew-symmetric matrix always satisfy $\beta_{2k-1} = \beta_{2k}$ for all $k \leq n/2$, and $\beta_n = 0$ if n is odd; and the matrix can be written as $V^t(\sum_{k\leq n/2} \beta_{2k}(E_{2k-1,2k} - E_{2k,2k-1}))V$ for some $V \in \mathcal{U}_n$.

Suppose α_j's and β_j's satisfy the conditions in (2.1) and (2.2). Let

$$\tilde{A} = \operatorname{diag}(\alpha_n, \ldots, \alpha_1) \in S_n(\mathbb{R}), \quad \tilde{B} = \sum_{j\leq n/2} \beta_{2j}(E_{2j-1,2j} - E_{2j,2j-1}) \in K_n(\mathbb{R}),$$

and

$$\Delta(\alpha, \beta) := \{\det(U^t\tilde{A}U + V^t\tilde{B}V) : U, V \in \mathcal{U}_n\}.$$

(Note that the diagonal entries of $\tilde{A}$ are arranged in ascending order instead of descending order so that it is easier to state our main results: Theorems 2.1 and 3.4.) We have the following theorem.

THEOREM 2.1. *The set $\Delta(\alpha, \beta) \subseteq \mathbb{C}$ is an annulus centered at the origin with outer radius equal to $\det(\tilde{A} + \tilde{B})$ and inner radius equal to*

$$\begin{cases} 0 & \text{if } [\alpha_{n-1}\alpha_n, \alpha_1\alpha_2] \cap [\beta_n^2, \beta_1^2] \neq \phi, \\ |\det(\tilde{A} + i\tilde{B})| & \text{otherwise.} \end{cases}$$

The proof of Theorem 2.1 is divided into several lemmas. We shall let

$$\tilde{\Delta}(\alpha, \beta) := \{U^t\tilde{A}U + V^t\tilde{B}V : U, V \in \mathcal{U}_n\}.$$

LEMMA 2.2. *The set $\Delta(\alpha, \beta) \subseteq \mathbb{C}$ is an annulus centered at the origin.*

Proof. Every $z \in \Delta(\alpha, \beta)$ can be regarded as the image of $(U, V) \in \mathcal{U}_n \times \mathcal{U}_n$ under the continuous mapping $f(U, V) = \det(U^t\tilde{A}U + V^t\tilde{B}V)$. Since $\mathcal{U}_n$ is path connected, it follows that $\Delta(\alpha, \beta)$ is path connected. Furthermore, if $z = f(U, V) \in \Delta(\alpha, \beta)$, then for any $\mu \in \mathbb{C}$ with $|\mu| = 1$ we can find a diagonal unitary matrix D such that $\det(D^2) = \mu$ so that $\mu z = f(DU, DV) \in \Delta(\alpha, \beta)$. Thus, $\mu\Delta(\alpha, \beta) = \Delta(\alpha, \beta)$ for any $\mu \in \mathbb{C}$ with $|\mu| = 1$. Combining the above arguments, we get the conclusion. ∎

By Lemma 2.2, $\Delta(\alpha, \beta) \subseteq \mathbb{C}$ is an annulus. It remains to determine the inner and outer radii of the annulus. We first study the case when the inner radius is 0, i.e., $0 \in \Delta(\alpha, \beta)$, or, equivalently, $\tilde{\Delta}(\alpha, \beta)$ contains a singular matrix.

LEMMA 2.3. *Every matrix in $\tilde{\Delta}(\alpha, \beta)$ is invertible, i.e., $0 \notin \Delta(\alpha, \beta)$ if and only if*

$$[\alpha_{n-1}\alpha_n, \alpha_1\alpha_2] \cap [\beta_n^2, \beta_1^2] = \phi.$$

Proof. First, we show that $0 \in \Delta(\alpha, \beta)$ if $[\alpha_{n-1}\alpha_n, \alpha_1\alpha_2] \cap [\beta_n^2, \beta_1^2] \neq \phi$. If n is odd, then $\alpha_1\alpha_2 > 0 = \beta_n^2$ and $\beta_1^2 \geq \alpha_{n-1}\alpha_n$. Consider $A(r) = U(r)A_0U(r)^t \oplus A_1$, where $A_0 = \operatorname{diag}(\alpha_n, -\alpha_{n-1}, \alpha_{n-2})$, $A_1 = \operatorname{diag}(\alpha_{n-3}, \ldots, \alpha_1)$, and

$$U(r) = [1] \oplus \begin{pmatrix} \cos r & \sin r \\ -\sin r & \cos r \end{pmatrix}.$$

Let $B = B_0 \oplus B_1$, where

$$B_0 = \begin{pmatrix} 0 & \beta_2 \\ -\beta_2 & 0 \end{pmatrix} \oplus [0], \qquad B_1 = \begin{pmatrix} 0 & \beta_4 \\ -\beta_4 & 0 \end{pmatrix} \oplus \cdots \oplus \begin{pmatrix} 0 & \beta_{n-1} \\ -\beta_{n-1} & 0 \end{pmatrix}.$$

One easily shows that the continuous real-valued function $g(r) = \det(A(r) + B) \in \Delta(\alpha, \beta)$ satisfies $g(0) \geq 0$ and $g(\pi/2) \leq 0$. Thus there exists $r_0 \in [0, \pi/2]$ such that $g(r_0) = 0$.

Now suppose n is even. Then $\beta_1^2 \geq \alpha_{n-1}\alpha_n$ and $\alpha_1\alpha_2 \geq \beta_n^2$. For $n = 2$, we have

$$0 = \det \begin{pmatrix} -\alpha_1 & \beta_2 \\ -\beta_2 & \alpha_2 \end{pmatrix}.$$

If $n \geq 4$, let $A(r) = U(r)A_0U(r)^t \oplus A_1$, where $A_0 = \operatorname{diag}(-\alpha_n, \alpha_{n-1}, -\alpha_2, \alpha_1)$, $A_1 = \operatorname{diag}(\alpha_3, \ldots, \alpha_{n-2})$, and

$$U(r) = [1] \oplus \begin{pmatrix} \cos r & \sin r \\ -\sin r & \cos r \end{pmatrix} \oplus [1].$$

Let

$$B = \begin{pmatrix} 0 & \beta_2 \\ -\beta_2 & 0 \end{pmatrix} \oplus \begin{pmatrix} 0 & \beta_n \\ -\beta_n & 0 \end{pmatrix} \oplus \begin{pmatrix} 0 & \beta_4 \\ -\beta_4 & 0 \end{pmatrix} \oplus \cdots \oplus \begin{pmatrix} 0 & \beta_{n-2} \\ -\beta_{n-2} & 0 \end{pmatrix}.$$

Then the real-valued function $h(r) = \det(A(r) + B) \in \Delta(\alpha, \beta)$ is continuous and satisfies $h(0) \leq 0$ and $h(\pi/2) \geq 0$, and so $h(r_0) = 0$ for some $r_0 \in [0, \pi/2]$.

Conversely, suppose $[\alpha_{n-1}\alpha_n, \alpha_1\alpha_2] \cap [\beta_n^2, \beta_1^2] = \phi$. Let $z = \det(U^t\tilde{A}U + V^t\tilde{B}V) \in \Delta(\alpha, \beta)$, where $U, V \in \mathcal{U}_n$. To prove that $z \neq 0$, we consider two cases (*not* in terms of the parity of n).

If $\alpha_{n-1}\alpha_n > \beta_1^2$, then $\tilde{A}^{-1}$ exists. Let $D = \tilde{A}^{-1/2}$. Then $Y = D(U^t)^*V^t\tilde{B}VU^*D$ is skew-symmetric, and we have (e.g., see [TT] and [MO, pp. 246–248])

$$s_1(Y)^2 = s_1(Y)s_2(Y) \leq s_1(D)^2 s_2(D)^2 s_1(\tilde{B})^2 \leq (\alpha_{n-1}\alpha_n)^{-1}\beta_1^2 < 1.$$

As a result, we have (e.g., see [MO, Chapter 9, G.1.e]) $s_n(I+Y) \geq s_n(I) - s_1(Y) > 0$ and $z = \det(U^tU)\det(\tilde{A})\det(I+Y) \neq 0$.

Suppose $\beta_n^2 > \alpha_1\alpha_2$. If $\tilde{A} = 0$, then $z = \det(V^t\tilde{B}V) \neq 0$. If $\tilde{A}$ has rank k, set $D = \{\gamma I_{n-k} \oplus \operatorname{diag}(\alpha_k, \ldots, \alpha_1)\}^{-1/2}$, where $\gamma = \alpha_k$ if $k > 1$, and γ satisfies $\min\{\beta_n^2/\alpha_1, \alpha_1\} > \gamma > 0$ if $k = 1$. Then $Y = D(U^t)^*V^t\tilde{B}VU^*D$ is skew-symmetric, and (see [TT] and [MO, pp. 246–248])

$$s_n(Y)^2 = s_n(Y)s_{n-1}(Y) \geq s_n(D)^2 s_{n-1}(D)^2 s_n(\tilde{B})^2 = \begin{cases} (\alpha_1\alpha_2)^{-1}\beta_n^2 > 1 & \text{if } k > 1, \\ (\alpha_1\gamma)^{-1}\beta_n^2 > 1 & \text{if } k = 1. \end{cases}$$

As a result, we have (e.g., see [MO, Chapter 9, G.1.e]) $s_n(E+Y) \geq s_n(Y) - s_1(E) > 0$ and $z = \det(U^tU)\det(D^{-2})\det(E+Y) \neq 0$, where $E = 0_{n-k} \oplus I_k$. □

To complete the proof of Theorem 2.1, we need some more notation. Let $x = (x_1, \ldots, x_n)^t, y = (y_1, \ldots, y_n)^t \in \mathbb{R}^n$. We say that x is *weakly majorized* by y, denoted by $x \prec_w y$, if the sum of the k largest entries of x is not greater than that of y for $k = 1, \ldots, n$. If, in addition, $\sum_{j=1}^n x_j = \sum_{j=1}^n y_j$, then x is said to be *majorized* by y, denoted by $x \prec y$. The following result (see [MO, Chapter 2, section C and Chapter 3, section C]) is useful in our discussion.

LEMMA 2.4. *Let ϕ be be a continuous convex real-valued function defined on an open set of $\mathbb{R}$ containing the entries of $x = (x_1, \ldots, x_n)^t, y = (y_1, \ldots, y_n)^t \in \mathbb{R}^n$.*

(a) *If* $x \prec y$, *then* $\sum_{j=1}^n \phi(x_j) \le \sum_{j=1}^n \phi(y_j)$.

(b) *If* $x \prec_w y$ *and* ϕ *is increasing, then* $(\phi(x_1), \dots, \phi(x_n))^t \prec_w (\phi(y_1), \dots, \phi(y_n))^t$.

We are now ready to complete the proof of Theorem 2.1.

LEMMA 2.5. *The annulus* $\Delta(\alpha, \beta)$ *has outer radius* $\det(\tilde{A}+\tilde{B})$. *If* $[\alpha_{n-1}\alpha_n, \alpha_1\alpha_2] \cap [\beta_n^2, \beta_1^2] = \phi$, *then* $\Delta(\alpha, \beta)$ *has inner radius* $|\det(\tilde{A} + i\tilde{B})|$.

Proof. Since the set $\Delta(\alpha, \beta)$ varies continuously as α_j's and β_j's, we may focus our attention on the case when $\alpha_1 > \cdots > \alpha_n > 0$. For any $z \in \Delta(\alpha, \beta)$, there exist $U, V \in \mathcal{U}_n$ such that $z = \det(U^t\tilde{A}U + V^t\tilde{B}V)$. Let $M = \tilde{A}^{-1/2}(U^t)^*V^t\tilde{B}VU^*\tilde{A}^{-1/2}$. Then M is skew-symmetric and has eigenvalues $\pm\lambda_1, \pm\lambda_2, \dots$, with $|\lambda_1| \ge |\lambda_2| \ge \cdots$. (This follows from the fact that $\det(tI - M) = \det((tI - M)^t) = \det(tI + M) = (-1)^n\det((-t)I - M)$.) It follows that

$$|z|/|\det(\tilde{A})| = |\det(I + M)| = \left| \prod_{j \le n/2} (1 - \lambda_j^2) \right|.$$

Note (e.g., see [TT] and [MO, pp. 246–248]) that

$$\prod_{j=1}^p |\lambda_j|^2 \le \prod_{j=1}^p s_{2j}(M)^2 \le \prod_{j=1}^p \beta_{2j}^2/(\alpha_{n-2j+2}\alpha_{n-2j+1}) \quad \text{for all } p \le n/2.$$

Let r be the largest integer so that $|\lambda_r| > 0$, and let $\log(c_1, \dots, c_r) = (\log c_1, \dots, \log c_r)$. Then

$$\begin{aligned} &\log(|\lambda_1|^2, \dots, |\lambda_r|^2) \ \prec_w \ \log(s_2(M)^2, \dots, s_{2r}(M)^2) \\ &\prec_w \ \log\left((\beta_1\beta_2)/(\alpha_n\alpha_{n-1}), \dots, (\beta_{2r-1}\beta_{2r})/(\alpha_{n-2r+2}\alpha_{n-2r+1})\right). \end{aligned} \tag{2.3}$$

To determine the outer radius of $\Delta(\alpha, \beta)$, note that the function $\phi(x) = \log(1 + e^x)$ is convex increasing on $\mathbb{R}$. Thus, we can apply Lemma 2.4 to the vectors in (2.3) and conclude that

$$\begin{aligned} |\det(I + M)| = \left| \prod_{j=1}^r (1 - \lambda_j^2) \right| &\le \prod_{j=1}^r (1 + |\lambda_j|^2) \le \prod_{j=1}^r (1 + s_{2j}(M)^2) \\ &\le \prod_{j=1}^r \left(1 + \frac{\beta_{2j}^2}{\alpha_{n-2j+2}\alpha_{n-2j+1}}\right) \le \frac{\det(\tilde{A} + \tilde{B})}{|\det(\tilde{A})|}. \end{aligned}$$

Since $\det(\tilde{A} + \tilde{B}) \in \Delta(\alpha, \beta)$, it must be the outer radius of $\Delta(\alpha, \beta)$.

Next suppose $[\alpha_{n-1}\alpha_n, \alpha_1\alpha_2] \cap [\beta_n^2, \beta_1^2] = \phi$, and consider the inner radius of $\Delta(\alpha, \beta)$. By Lemma 2.3, there are two possibilities.

First, assume $\beta_1^2 < \alpha_{n-1}\alpha_n$. Then $|\lambda_1|^2 \le |s_1(M)|^2 \le \beta_1^2/(\alpha_{n-1}\alpha_n) < 1$. Since the function $\phi(x) = -\log(1 - e^x)$ is convex increasing for $x < 0$, we can apply Lemma 2.4 to the vectors in (2.3) and conclude that

$$\begin{aligned} |\det(I + M)| = \left| \prod_{j=1}^r (1 - \lambda_j^2) \right| &\ge \prod_{j=1}^r (1 - |\lambda_j|^2) \ge \prod_{j=1}^r (1 - s_{2j}(M)^2) \\ &\ge \prod_{j=1}^r \left(1 - \frac{\beta_{2j}^2}{\alpha_{n-2j+2}\alpha_{n-2j+1}}\right) \ge \frac{\det(\tilde{A} + i\tilde{B})}{|\det(\tilde{A})|}. \end{aligned}$$

Thus $\det(\tilde{A}+i\tilde{B}) \in \Delta(\alpha,\beta)$ is the inner radius of $\Delta(\alpha,\beta)$ in this case.

Next assume $\beta_n^2 > \alpha_1\alpha_2$. Then $|\lambda_{n/2}|^2 \geq |s_n(M)|^2 \geq \beta_n^2/(\alpha_1\alpha_2) > 1$, and hence M is invertible. As a result, we have $r = n/2$ in (2.3) and the weak majorizations are actually majorizations. Since the function $\phi(x) = -\log(e^x - 1)$ is convex for $x > 0$, we can apply Lemma 2.4 to the vectors in (2.3) and conclude that

$$|\det(I+M)| = \left|\prod_{j=1}^{r}(1-\lambda_j^2)\right| \geq \prod_{j=1}^{r}(|\lambda_j|^2-1) \geq \prod_{j=1}^{r}(s_{2j}(M)^2-1)$$
$$\geq \prod_{j=1}^{r}\left(\frac{\beta_{2j}^2}{\alpha_{n-2j+2}\alpha_{n-2j+1}}-1\right) = \frac{\det(i\tilde{A}+\tilde{B})}{|\det(\tilde{A})|}.$$

Thus $\det(i\tilde{A}+\tilde{B}) \in \Delta(\alpha,\beta)$ is the inner radius of $\Delta(\alpha,\beta)$ in this case. □

Our technique can be used to estimate the principal minors of a given square matrix X. In particular, we can determine the set $\Delta_k(\alpha,\beta)$ of all $k \times k$ principal minors of $X \in \tilde{\Delta}(\alpha,\beta)$. Since $X \in \tilde{\Delta}(\alpha,\beta)$ if and only if $PXP^t \in \tilde{\Delta}(\alpha,\beta)$ for any permutation matrix P, we can focus our attention on the determinant of the leading $k \times k$ principal submatrix $X[k]$ of $X \in \tilde{\Delta}(\alpha,\beta)$. We have the following result (cf. [T2] and [Ta] for the case when $k = 1$).

THEOREM 2.6. *Suppose* $1 \leq k < n$. *Then*

$$\Delta_k(\alpha,\beta) := \{\det(X[k]) : X \in \tilde{\Delta}(\alpha,\beta)\}$$

is a circular disk centered at the origin with radius $\det(\tilde{A}_k+\tilde{B}_k)$, *where*

$$\tilde{A}_k = \operatorname{diag}(\alpha_k,\ldots,\alpha_1),\ \tilde{B}_k = \sum_{j\leq k/2}\beta_{2j}(E_{2j-1,2j}-E_{2j,2j-1}) \in M_k(\mathbb{C}).$$

Proof. Applying the arguments in the proof of Lemma 2.2 to the continuous function $f_k(U,V) = \det((U^t\tilde{A}U+V^t\tilde{B}V)[k])$ defined on $\mathcal{U}_n \times \mathcal{U}_n$, we see that $\Delta_k(\alpha,\beta)$ is an annulus in $\mathbb{C}$ centered at the origin.

Next, we show that $0 \in \Delta_k(\alpha,\beta)$ to conclude that $\Delta_k(\alpha,\beta)$ is a circular disk. To this end, consider $U \in \mathcal{U}_n$ obtained from I by replacing the $(1,1),(1,n),(n,1),(n,n)$ entries with $i\cos r$, $i\sin r$, $\sin r$, $-\cos r$ for some suitable $r \in \mathbb{R}$ such that the leading $k\times k$ principal submatrix of $U^t(\sum_{j=1}^n \alpha_j E_{jj})U$ is of the form $A_k = \operatorname{diag}(0,\alpha_2,\ldots,\alpha_k)$ $\in M_k(\mathbb{C})$. Also, there exists a permutation matrix P such that the leading $k \times k$ principal submatrix of $P^t\tilde{B}P$ is of the form $B_k = 0_1 \oplus B_0$ with $B_0 = -B_0^t \in M_{k-1}$. Then $\det(A_k+B_k) = 0 \in \Delta_k(\alpha,\beta)$.

To determine the (outer) radius of $\Delta_k(\alpha,\beta)$, let $Z = Z_1+Z_2$ with $Z_1 = (U^t\tilde{A}U)[k]$ and $Z_2 = (V^t\tilde{B}V)[k]$ such that $|\det(Z)|$ attains the radius of $\Delta_k(\alpha,\beta)$. By Theorem 2.1, we may assume that

$$Z_1 = \operatorname{diag}(\mu_k,\ldots,\mu_1),\ Z_2 = \sum_{j\leq k/2}\nu_{2j}(E_{2j-1,2j}-E_{2j,2j-1}) \in M_k(\mathbb{C}),$$

where $\mu_j = s_j(Z_1)$ and $\nu_{2j} = s_{2j}(Z_2)$. Note that (e.g., see [T1]) $\mu_j \leq \alpha_j$ for $j = 1,\ldots,k$, and $\nu_{2j} \leq \beta_{2j}$ for all $j \leq k/2$. It follows that $|\det(Z)| = \det(\tilde{A}_k+\tilde{B}_k)$. □

3. Results on real matrices. The results on real matrices are more complicated. One of the reasons is that the topological structure of $\mathcal{O}_n$ is not as nice as that of $\mathcal{U}_n$, and hence the set of real symmetric matrices with prescribed singular values is more complicated. In fact, we have the following description of the set of real symmetric and the set of real skew-symmetric matrices with prescribed singular values (e.g., see [HJ, Chapter 3] and [Y]).

(3.1) every real symmetric matrix with singular values $\alpha_1 \geq \cdots \geq \alpha_n$ can be written as $U^t(\sum_{i=1}^n \varepsilon_i \alpha_i E_{ii})U$ for some $U \in \mathcal{O}_n$ and $(\varepsilon_1, \ldots, \varepsilon_n) \in \{-1, 1\}^n$;

(3.2) the singular values $\beta_1 \geq \cdots \geq \beta_n$ of a real skew-symmetric matrix always satisfy $\beta_{2k-1} = \beta_{2k}$ for all $k \leq n/2$, and $\beta_n = 0$ if n is odd; and the matrix can be written as $V^t(\sum_{k \leq n/2} \beta_{2k}(E_{2k-1,2k} - E_{2k,2k-1}))V$ for some $V \in \mathcal{O}_n$.

Suppose α_j's and β_j's satisfy the conditions in (3.1) and (3.2). We continue to use the notation $\tilde{A}$, $\tilde{B}$, $\Delta(\alpha, \beta)$, and $\tilde{\Delta}(\alpha, \beta)$ as in the complex case. Let

$$\Delta^{\mathbb{R}}(\alpha, \beta) := \{\det(U^t E \tilde{A} U + V^t \tilde{B} V) : E, U, V \in \mathcal{O}_n, \text{ where } E \text{ is in diagonal form}\}.$$

Clearly, we have

$$\Delta^{\mathbb{R}}(\alpha, \beta) = \{\det(X) : X \in \tilde{\Delta}(\alpha, \beta) \cap M_n(\mathbb{R})\} \subseteq \Delta(\alpha, \beta) \cap \mathbb{R}.$$

We shall show that the set inclusion is an equality if n is odd (cf. Theorem 3.2), and the set inclusion may be proper if n is even (cf. Proposition 3.3 and Theorem 3.4). We first study the following subsets of $\Delta^{\mathbb{R}}(\alpha, \beta)$:

$\Delta^{\varepsilon}(\alpha, \beta) := \{\det(A + B) : A \in S_n(\mathbb{R})$ has eigenvalues $\varepsilon_1\alpha_1, \ldots, \varepsilon_n\alpha_n$, $B \in K_n(\mathbb{R})$ has singular values $\beta_1, \ldots, \beta_n\}, (\varepsilon_1, \ldots, \varepsilon_n) \in \{-1, 1\}^n$.

$\Delta^{\iota}(\alpha, \beta) := \Delta^{\varepsilon}(\alpha, \beta)$ with $\varepsilon_j = 1$ for all $j = 1, \ldots, n$.

$$\Delta^{+}(\alpha, \beta) := \bigcup \Big\{\Delta^{\varepsilon}(\alpha, \beta) : (\varepsilon_1, \ldots, \varepsilon_n) \in \{-1, 1\}^n \text{ with } \textstyle\prod_{j=1}^n \varepsilon_j = 1\Big\}.$$

$$\Delta^{-}(\alpha, \beta) := \bigcup \Big\{\Delta^{\varepsilon}(\alpha, \beta) : (\varepsilon_1, \ldots, \varepsilon_n) \in \{-1, 1\}^n \text{ with } \textstyle\prod_{j=1}^n \varepsilon_j = -1\Big\}.$$

Evidently,

$$\Delta^{\mathbb{R}}(\alpha, \beta) = \Delta^{+}(\alpha, \beta) \cup \Delta^{-}(\alpha, \beta),$$

and by Theorem 2.1 we see that

$$z_0 = \det(\tilde{A} + \tilde{B})$$

is the right endpoint of $\Delta^{\iota}(\alpha, \beta)$, $\Delta^{+}(\alpha, \beta)$, and $\Delta^{\mathbb{R}}(\alpha, \beta)$. Furthermore, we have the following result.

THEOREM 3.1. *The sets $\Delta^{+}(\alpha, \beta)$, $\Delta^{-}(\alpha, \beta)$, and $\Delta^{\varepsilon}(\alpha, \beta)$ for any $(\varepsilon_1, \ldots, \varepsilon_n) \in \{-1, 1\}^n$ are all closed intervals in $\mathbb{R}$. Moreover, $\Delta^{+}(\alpha, \beta) = -\Delta^{-}(\alpha, \beta)$ if n is odd.*

Proof. Suppose $(\varepsilon_1, \ldots, \varepsilon_n) \in \{-1, 1\}^n$. Set $E = \operatorname{diag}(\varepsilon_1, \ldots, \varepsilon_n)$. For every $z = \det(U^t E \tilde{A} U + V^t \tilde{B} V)$ with $U, V \in \mathcal{O}_n$, we can let $D = \operatorname{diag}(\pm 1, 1, \ldots, 1)$ such that $\det(VU^t D) = 1$ and $z = \det(E\tilde{A} + UV^t \tilde{B} V U^t) = \det(E\tilde{A} + DUV^t \tilde{B} V U^t D)$. Thus

$$\begin{aligned}\Delta^{\varepsilon}(\alpha, \beta) &= \{\det(U^t E \tilde{A} U + V^t \tilde{B} V) : U, V \in \mathcal{O}_n\} \\ &= \{\det(E\tilde{A} + W^t \tilde{B} W) : W \in \mathcal{O}_n, \det(W) = 1\}.\end{aligned}$$

As a result, the set $\Delta^{\varepsilon}(\alpha,\beta)$ can be viewed as the image of the compact connected set $\mathcal{O}_n^+ = \{W \in \mathcal{O}_n : \det(W) = 1\}$ under the continuous function $f(W) = \det(E\tilde{A} + W^t\tilde{B}W)$ and must be a closed interval in $\mathbb{R}$.

To prove that $\Delta^+(\alpha,\beta)$ is a closed interval, we show that $\Delta^{\varepsilon}(\alpha,\beta) \cap \Delta^{\iota}(\alpha,\beta) \neq \phi$ for any $(\varepsilon_1,\ldots,\varepsilon_n) \in \{-1,1\}^n$ with $\prod_{j=1}^n \varepsilon_j = 1$. Suppose k of the ε_j's are -1, where $k \geq 0$ is even. Then there is a permutation matrix P such that $P^tEP = -I_k \oplus I_{n-k}$, where $E = \operatorname{diag}(\varepsilon_1,\ldots,\varepsilon_n)$. It follows that the first k entries of $P^tE\tilde{A}P$ are negative, and $z = \det(P^tE\tilde{A}P+\tilde{B}) = \det(P^t\tilde{A}P+\tilde{B}) \in \Delta^{\varepsilon}(\alpha,\beta)\cap\Delta^{\iota}(\alpha,\beta)$. Similarly, one can prove that if $\eta = (1,\ldots,1,-1)$, then $\Delta^{\varepsilon}(\alpha,\beta) \cap \Delta^{\eta}(\alpha,\beta) \neq \phi$ for any $(\varepsilon_1,\ldots,\varepsilon_n) \in \{-1,1\}^n$ with $\prod_{j=1}^n \varepsilon_j = -1$. Hence $\Delta^-(\alpha,\beta)$ is also a closed interval.

Finally, if n is odd, then $z = \det(A+B) \in \Delta^+(\alpha,\beta)$ if and only if $-z = \det(-A-B) \in \Delta^-(\alpha,\beta)$. The last assertion of the theorem follows. □

As a consequence of Theorem 3.1, we see that $\Delta^{\mathbb{R}}(\alpha,\beta)$ is the union of the two closed intervals $\Delta^+(\alpha,\beta)$ and $\Delta^-(\alpha,\beta)$. We obtain more information for $\Delta^{\mathbb{R}}(\alpha,\beta)$ in the following.

THEOREM 3.2. *Suppose $n > 1$ is an odd integer. Then $\Delta^{\mathbb{R}}(\alpha,\beta) = \Delta(\alpha,\beta) \cap \mathbb{R}$.*

Proof. Recall that $z_0 = \det(\tilde{A}+\tilde{B})$ is the right endpoint of $\Delta^+(\alpha,\beta)$. By Theorem 3.1, $-z_0$ is the left endpoint of $\Delta^-(\alpha,\beta)$.

If $[\alpha_{n-1}\alpha_n, \alpha_1\alpha_2] \cap [\beta_n^2, \beta_1^2] \neq \phi$, the proof of Lemma 2.3 actually shows that $0 \in \Delta^-(\alpha,\beta)$, and hence $[-z_0, 0] \subseteq \Delta^-(\alpha,\beta)$. Thus, $[0, z_0] \subseteq \Delta^+(\alpha,\beta)$, and $\Delta^{\mathbb{R}}(\alpha,\beta) = [-z_0, z_0] = \Delta(\alpha,\beta) \cap \mathbb{R}$.

If $[\alpha_{n-1}\alpha_n, \alpha_1\alpha_2]\cap[\beta_n^2, \beta_1^2] = \phi$, then $\alpha_{n-1}\alpha_n > \beta_1^2$. Let $A = \sum_{j=1}^n (-1)^{j-1}\alpha_{n-j+1}E_{jj}$. If $n = 4k+1$ for some nonnegative integer k, then $\det(A+\tilde{B}) \in \Delta^+(\alpha,\beta)$ is positive. If $n = 4k+3$ for some nonnegative integer k, then $\det(A+\tilde{B}) \in \Delta^-(\alpha,\beta)$ is negative. In both cases, $z_1 = |\det(A+\tilde{B})| \in \Delta^+(\alpha,\beta)$, and hence $[z_1, z_0] \subseteq \Delta^+(\alpha,\beta)$. Note that $z_1 = |\det(\tilde{A}+i\tilde{B})|$. Thus, $\Delta(\alpha,\beta) \cap \mathbb{R} = [-z_0, -z_1] \cup [z_1, z_0] \subseteq \Delta^{\mathbb{R}}(\alpha,\beta)$. □

Next, we turn to the case when n is even.

PROPOSITION 3.3. *Suppose $n = 2$. Then $\Delta^{\mathbb{R}}(\alpha,\beta) = \{\beta_1^2 - \alpha_1\alpha_2, \beta_1^2 + \alpha_1\alpha_2\}$.*

Proof. The proof is by direct verification. □

By Proposition 3.3, we see that for even n it is hopeless to get $\Delta^{\mathbb{R}}(\alpha,\beta) = \Delta(\alpha,\beta) \cap \mathbb{R}$ in general. Nevertheless, we have the following theorem.

THEOREM 3.4. *Suppose $n \geq 4$ is an even integer. Let $z_0 = \det(\tilde{A}+\tilde{B})$ and $z_1 = \det(E_1\tilde{A}+\tilde{B})$ with $E_1 = \sum_{j=1}^n (-1)^j E_{jj}$. Then one of the following holds.*

(a) *If $[\alpha_{n-1}\alpha_n, \alpha_1\alpha_2] \cap [\beta_n^2, \beta_1^2] \neq \phi$, then $\Delta^{\mathbb{R}}(\alpha,\beta) = [a, z_0]$ for some $a \leq 0$, where $a = 0$ if and only if $\alpha_1\alpha_2 = \beta_n^2$ or $z_0 = 0$.*

(b) *If $\alpha_1\alpha_2 < \beta_n^2$, then $\Delta^{\mathbb{R}}(\alpha,\beta) = [z_1, z_0] = \Delta(\alpha,\beta) \cap (0,\infty)$.*

(c) *If $\alpha_n\alpha_{n-1} > \beta_1^2$, then $\Delta^-(\alpha,\beta) \subseteq (-\infty, 0)$ and $\Delta^+(\alpha,\beta) \subseteq (0,\infty)$ have no intersection. Furthermore,*

$$\Delta^{\mathbb{R}}(\alpha,\beta) = \Delta^-(\alpha,\beta) \cup \Delta^+(\alpha,\beta) = \begin{cases} [b_1, b_2] \cup [z_1, z_0] & \text{if } n = 4k, \\ [c_1, z_1] \cup [c_2, z_0] & \text{if } n = 4k+2. \end{cases}$$

Proof. Let $z_2 = \det(E_2\tilde{A}+\tilde{B})$, $z_3 = \det(E_3\tilde{A}+\tilde{B}) \in \Delta^-(\alpha,\beta)$, where $E_2 = [-1] \oplus I_{n-1}$ and $E_3 = I_{n-1} \oplus [-1]$.

(a) Suppose $[\alpha_{n-1}\alpha_n, \alpha_1\alpha_2]\cap[\beta_n^2, \beta_1^2] \neq \phi$. One can use the arguments in the proof of Lemma 2.3 to show that $0 \in \Delta^{\varepsilon}(\alpha,\beta) \subseteq \Delta^+(\alpha,\beta)$, where $\varepsilon_2 = \varepsilon_n = -1$ and $\varepsilon_j = 1$ for all other j. Moreover, $z_2 \in \Delta^-(\alpha,\beta)$ is positive. Thus $\Delta^+(\alpha,\beta) \cap \Delta^-(\alpha,\beta) \neq \phi$ and $\Delta^{\mathbb{R}}(\alpha,\beta) = [a, z_0]$ with $a \leq z_3 \leq 0$. If $a = 0$, then $z_3 = 0$ and so $\beta_n^2 = \alpha_1\alpha_2$

or $z_0 = 0$. Conversely, if $\beta_n^2 = \alpha_1\alpha_2$, then we have $\Delta^{\mathbb{R}}(\alpha,\beta) = [z_1, z_0]$ $(= [0, z_0])$ by applying a continuity argument to (b). Also, if $z_0 = 0$, then $\Delta^{\mathbb{R}}(\alpha,\beta) = \{0\}$ by Theorem 2.1.

(b) Suppose $\alpha_1\alpha_2 < \beta_n^2$. Then $z_2 \in \Delta^-(\alpha,\beta)$ is positive. Let $z_4 = \det(E_4\tilde{A} + \tilde{B})$, where $E_4 = \operatorname{diag}(-1,1,-1) \oplus I_{n-3}$. Then $z_4 \in \Delta^+(\alpha,\beta)$ and $z_4 \le z_2$. Thus $\Delta^+(\alpha,\beta) \cap \Delta^-(\alpha,\beta) \neq \phi$ and thus $\Delta^{\mathbb{R}}(\alpha,\beta) = [r, z_0]$. Since $\Delta^{\mathbb{R}}(\alpha,\beta) \subseteq \Delta(\alpha,\beta) \cap \mathbb{R} = [-z_0, -z_1] \cup [z_1, z_0]$, and since $z_1, z_0 \in \Delta^{\mathbb{R}}(\alpha,\beta)$, we have $\Delta^{\mathbb{R}}(\alpha,\beta) = [z_1, z_0]$.

(c) Suppose $\alpha_n\alpha_{n-1} > \beta_1^2$. Then $z_3 \in \Delta^-(\alpha,\beta)$ is negative. Recall that we always have $z_0 \in \Delta^+(\alpha,\beta)$. Since $\Delta(\alpha,\beta) \cap \mathbb{R}$ is disconnected and is symmetric about the origin, we conclude that $\Delta^-(\alpha,\beta) \subseteq (-\infty, 0)$ and $\Delta^+(\alpha,\beta) \subseteq (0,\infty)$ have no intersection. Furthermore, if $n = 4k$, then $z_1 > 0$ attains the inner radius of $\Delta(\alpha,\beta)$. Thus $\Delta^+(\alpha,\beta) = [z_1, z_0]$. If $n = 4k+2$, then $z_1 < 0$ and $|z_1|$ attains the inner radius of $\Delta(\alpha,\beta)$. Thus z_1 is the right endpoint of $\Delta^-(\alpha,\beta)$. □

Similar to the complex case, we can get estimates for the principal minors of matrices of the form $X = U^t\tilde{A}U + V^t\tilde{B}V$, where $U, V \in \mathcal{O}_n$. We omit the details.

4. Remarks and open problems. In [LM] (see also [Mi]), it is shown that there exist $A, B \in M_n(\mathbb{F})$ with singular values $\alpha_1 \ge \cdots \ge \alpha_n$ and $\beta_1 \ge \cdots \ge \beta_n$ such that $\det(A+B) = z$ if and only if

$$\prod_{j=1}^{n}(\alpha_j + \beta_{n-j+1}) \ge |z| \ge \begin{cases} 0 & \text{if } [\alpha_n, \alpha_1] \cap [\beta_n, \beta_1] \neq \phi, \\ \left|\prod_{j=1}^{n}(\alpha_j - \beta_{n-j+1})\right| & \text{otherwise.} \end{cases}$$

Note that in this result the containment region of $\det(A+B)$ for the real case is simply the intersection of the containment region for the complex case and the real line. Also note that since there are more restrictions on the structure of A and B in our study, we have better estimates (smaller containment region) for $\det(A+B)$.

We were not able to determine the values a, b_1, b_2, c_1, c_2 in Theorem 3.4. It would be nice to find formulae for these quantities. If this is done, then there will be complete descriptions of $\Delta^{\mathbb{R}}(\alpha,\beta)$, $\Delta^+(\alpha,\beta)$, and $\Delta^-(\alpha,\beta)$.

It is also interesting to determine the intervals $\Delta^{\varepsilon}(\alpha,\beta)$. Note that $\Delta^{\varepsilon}(\alpha,\beta)$ can be viewed as the set of numbers of the form $\det(A+B)$, where $A \in S_n(\mathbb{R})$ has eigenvalues $\varepsilon_j\alpha_j$'s and $B \in K_n(\mathbb{R})$ has eigenvalues $\pm i\beta_{2j}$'s. One can consider the following analogous complex problem: *Study the set of complex numbers of the form* $\det(H+iK)$*, where* H *and* K *are complex Hermitian matrices with prescribed eigenvalues.*

This is actually a special case of the Marcus–Oliveira conjecture (see [Ma] and [O]):

Let $C = \operatorname{diag}(c_1,\ldots,c_n)$, $D = \operatorname{diag}(d_1,\ldots,d_n) \in M_n(\mathbb{C})$. *Suppose*

$$\Gamma(c,d) = \{\det(U^*CU + V^*DV) : U, V \in \mathcal{U}_n\}.$$

Then $\Gamma(c,d)$ *is a subset of the convex hull of the set*

$$P(c,d) = \{\det(C + P^tDP) : P \textit{ is a permutation matrix}\}.$$

This conjecture has only been confirmed in some special cases (see [B] and its references). Note that if $\alpha E = (\alpha_n,\ldots,\alpha_1)E$ with $E = \operatorname{diag}(\varepsilon_1,\ldots,\varepsilon_n) \in \mathcal{O}_n$, and $(d_1,\ldots,d_n) = (i\beta_2, -i\beta_2, i\beta_4, -i\beta_4, \ldots)$, where $d_n = 0$ if n is odd, then

$$\Gamma(\alpha E, d) = \{\det(U^*E\tilde{A}U + V^*\tilde{B}V) : U, V \in \mathcal{U}_n\}.$$

Thus we have

$$\Delta^{\varepsilon}(\alpha,\beta) \subseteq \Gamma(\alpha E, d) \cap \mathbb{R}$$

and hence

$$\Delta^{\mathbb{R}}(\alpha,\beta) \subseteq \bigcup_{\varepsilon\in\{-1,1\}^n} \{\Gamma(\alpha E, d) \cap \mathbb{R}\}.$$

These provide other estimates for $\Delta^{\varepsilon}(\alpha,\beta)$ and $\Delta^{\mathbb{R}}(\alpha,\beta)$. It is worth mentioning that when $n = 2$ the Marcus–Oliveira conjecture is valid, and actually we have

$$\Delta^{\varepsilon}(\alpha,\beta) = \Gamma(\alpha E, d) \cap \mathbb{R}$$

(cf. [B] and Proposition 3.3).

Acknowledgment. The authors would like to thank the editor and the referees for many helpful comments.

REFERENCES

[B] N. Bebiano, *New developments on the Marcus-Oliveira conjecture*, Linear Algebra Appl., 197/198 (1994), pp. 793–803.

[C] A. L. Cauchy, *Mémoir sur les functions qui ne peuvent obtenir que deux valeurs égales et de signes contraires par suite de transpositions opérées entre les variables qu'elles renferment*, J. Ec. Polyt., 10 (1815), pp. 29–112.

[F] M. Fiedler, *Bounds for the determinant of the sum of Hermitian matrices*, Proc. Amer. Math. Soc., 30 (1971), pp. 27–31.

[HJ] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, London, 1985.

[LM] C. K. Li and R. Mathias, *Determinant of the sum of two matrices*, Bull. Austral. Math. Soc., 52 (1995), pp. 425–429.

[Ma] M. Marcus, *Derivations, Plücker relations and the numerical range*, Indiana Univ. Math. J., 22 (1973), pp. 1137–1149.

[MO] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*, Academic Press, New York, 1979.

[Mi] M. E. Miranda, *On the trace of the product and the determinant of the sum of complex matrices with prescribed singular values*, in Proc. Encuentro Internacional de Algebra Lineal y Aplicaciones, Gasteiz, Spain, 1984, pp. 326–333.

[Mu] T. Muir, *The Theory of Determinants in the Historical Order of Development,* Vol. 3, MacMillan, London, 1920.

[O] G. N. de Oliveira, *Normal matrices (research problem)*, Linear and Multilinear Algebra, 12 (1982), pp. 153–154.

[Ta] T. Y. Tam, *Note on a paper of Thompson: The congruence numerical range*, Linear and Multilinear Algebra, 17 (1985), pp. 107–115.

[T1] R. C. Thompson, *Principal submatrices* IX. *Interlacing inequalities for singular values*, Linear Algebra Appl., 5 (1972), pp. 1–12.

[T2] R. C. Thompson, *The congruence numerical range*, Linear and Multilinear Algebra, 8 (1980), pp. 197–206.

[TT] R. C. Thompson and S. Therianos, *On the singular values of matrix products*, I-III, Scripta Mathematica, 29 (1970), pp. 99–123.

[Y] D. C. Youla, *A normal form for a matrix under the unitary group*, Canad. J. Math., 13 (1961), pp. 694–704.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 83–103, January 1997

008

BOUNDS FOR THE COMPONENTWISE DISTANCE TO THE NEAREST SINGULAR MATRIX*

S. M. RUMP†

Abstract. The normwise distance of a matrix A to the nearest singular matrix is well known to be equal to $\|A\|/\mathrm{cond}(A)$ for norms subordinate to a vector norm. However, there is no hope for finding a similar formula or even a simple algorithm for computing the *componentwise distance* to the nearest singular matrix for general matrices. This is because Poljak and Rohn [*Math. Control Signals Systems*, 6 (1993), pp. 1–9] showed that this is an NP-hard problem.

Denote the minimum Bauer–Skeel condition number achievable by column scaling by κ. Demmel [*SIAM J. Matrix Anal. Appl.*, 13 (1992), pp. 10–19] showed that κ^{-1} is a *lower* bound for the componentwise distance to the nearest singular matrix. In our paper we prove that $2.4 \cdot n^{1.7} \cdot \kappa^{-1}$ is an *upper* bound. This extends and proves a conjecture by Demmel and Higham (in the cited paper by Demmel). We give an explicit set of examples showing that such an upper bound cannot be better than $n \cdot \kappa^{-1}$. Asymptotically, we show that $n^{1+\ln 2+\varepsilon} \cdot \kappa^{-1}$ is a valid upper bound.

Key words. componentwise distance, singular matrix, NP-hardness, optimal Bauer–Skeel condition number

AMS subject classifications. 65F35, 15A60

PII. S0895479895289170

Introduction. Let A be an n-by-n matrix and denote its smallest singular value by $\sigma_n(A)$. It is well known that the distance to the nearest singular matrix in the 2-norm or Frobenius norm is equal to $\sigma_n(A)$. More generally, for any consistent matrix norm $\|\cdot\|$ subordinate to a vector norm we have

$$\min\{\,\|\delta A\| \mid A+\delta A \text{ singular}\,\} = \frac{1}{\|A^{-1}\|} = \frac{\|A\|}{\mathrm{cond}(A)}. \tag{0.1}$$

An appropriate δA of rank one can be explicitly calculated (cf. [6], [14]). Such a perturbation does, in general, alter each component of A. In many practical applications, one is interested in leaving specific components such as system zeros unaltered, for example, if the matrix arises from some discretization scheme. More generally, this leads to the question of the componentwise distance to the nearest singular matrix. The componentwise distance may be weighted by some nonnegative matrix E. More precisely, we define

$$\sigma(A,E) := \min\{\,\alpha\in\mathbb{R} \mid A+\widetilde{E} \text{ singular where } |\widetilde{E}_{ij}| \le \alpha\cdot E_{ij} \text{ for all } i,j\,\}. \tag{0.2}$$

If no such α exists, we set $\sigma(A,E) := \infty$. For singular matrices, $\sigma(A,E) = 0$ for every weight matrix E. Specific values of E are $E = |A|$ for relative perturbations or $E = (\mathbf{1})_{nn}$ for absolute perturbations. Among others, the componentwise distance to the nearest singular matrix was discussed in [9], [12], [11], and [3]. In [9] we also find a first approach toward an estimation of the nearness to singularity in a norm not subordinate to a vector norm, namely, $\|A\| := \max_{i,j}|A_{ij}|$.

We cannot expect to find a formula or even a simple algorithm for calculating $\sigma(A,E)$. This is because Poljak and Rohn [8] proved that computation of $\sigma(A,E)$ is

* Received by the editors June 28, 1995; accepted for publication (in revised form) by N. J. Higham January 9, 1996.
http://www.siam.org/journals/simax/18-1/28917.html
† Technische Informatik III, TU Hamburg-Harburg, Eißendorfer Straße 38, 21071 Hamburg, Germany (rump@tu-harburg.d400.de).

NP-complete. For an outline of their proof see also [3]. Nevertheless, we may find bounds for $\sigma(A,E)$, and for classes of matrices even explicit formulas.

Another view of $\sigma(A,E)$ is the maximum value such that the interval matrix $[A-\alpha E,\ A+\alpha E]$ is nonsingular for $\alpha < \sigma(A,E)$. The interval matrix is defined as the set of all matrices $\widetilde{A}$ with $A_{ij}-\alpha E_{ij} \le \widetilde{A}_{ij} \le A_{ij}+\alpha E_{ij}$ for all i,j or, in short notation, $A-\alpha E \le \widetilde{A} \le A+\alpha E$. The interval matrix is called nonsingular if every matrix $\widetilde{A} \in [A-\alpha E,\ A+\alpha E]$ is nonsingular. In a very interesting paper [10], Rohn gave 13 necessary and sufficient criteria for $[A-E,\ A+E]$ to be nonsingular.

A thorough discussion of $\sigma(A,E)$ can be found in the very interesting paper [3]. Demmel [3] proved that $\sigma(A,E)$ is greater than or equal to the inverse of $\min \kappa(AD,\ ED)$, the minimum taken over all diagonal matrices D, where $\kappa(A,E) := \| \, |A^{-1}| \cdot E\|$ denotes the Bauer–Skeel condition number. For any p-norm, he proves

$$\min_D \ \kappa(AD,\ ED) = \rho(\, |A^{-1}| \cdot E),$$

extending a result by Bauer [1]. In other words, the minimum Bauer–Skeel condition number achievable by column scaling is equal to the inverse of $\rho(\, |A^{-1}| \cdot E)$. Demmel and Higham conjecture that $1/\rho(\, |A^{-1}| \cdot E)$ and $\sigma(A,E)$ are not too far apart. They conjecture for relative perturbations the existence of some constant $\gamma \in \mathbb{R}$, possibly depending on the dimension, with

$$\sigma(A, |A|\,) \le \frac{\gamma}{\rho(\, |A^{-1}| \cdot |A|\,)}. \tag{0.3}$$

In this paper, our main goal is to show the existence of such constants $\gamma(n)$ and to derive lower and upper bounds for $\gamma(n)$. First, we show that $\sigma(A,E) \ge \sigma_n(A)$ for $\|E\|_2 = 1$. A corresponding result for other norms is given in section 2. However, this bound can be arbitrarily weak. Following, we give some new bounds for $\sigma(A,E)$.

In section 4 a perturbation formula for determinants is stated, which is the key to proving an upper bound of $\gamma(n)$.

In section 5 we will prove $\gamma \ge n$. In section 6, for arbitrary weight matrices E, we prove

$$\frac{1}{\rho(\, |A^{-1}| \cdot E)} \le \sigma(A,E) \le \frac{\gamma(n)}{\rho(\, |A^{-1}| \cdot E)} \quad \text{with} \quad \gamma(n) = c \cdot n^{\alpha} \tag{0.4}$$

for $c = 2.4$ and $\alpha = 1.7$. Moreover, for $n \to \infty$ we show that for every $\varepsilon > 0$, α can be replaced by $1 + \ln 2 + \varepsilon$. In view of $\gamma \ge n$, we conjecture that $\gamma = n$.

In [3], Demmel gave reasons to be interested in the componentwise distance to the nearest singular matrix. In section 2, we add a lower and upper componentwise error bound for the solution of a linear system $Ax = b$ subject to componentwise perturbations of the matrix and the right-hand side. Such upper bounds are known in the literature and are valid for nonsingular A and $|\widetilde{A} - A| \le E$ with $\rho(|\, A^{-1}| \cdot E) < 1$. We derive a componentwise bound for the *minimum* perturbation of the solution subject to finite perturbations of A and b. Equation (0.4) shows that those estimates cover perturbation matrices $\widetilde{A}$ not too far from the next singular matrix.

The paper is organized as follows. In section 1 we introduce the used notation. In section 2 there follows a componentwise lower and upper perturbation bound for finite componentwise perturbations of a linear system. In section 3, lower bounds on $\sigma(A,E)$ are given. For orthogonal matrices we show that γ (see (0.4)) is at least of the order of $\sqrt{n}$.

In section 4, a Sherman–Morrison–Woodbury-like perturbation theorem for determinants is given. In fact, this is an *equality* for finite perturbations of a matrix. In section 5 we derive upper bounds on $\sigma(A,E)$. For E of rank one, such as for absolute perturbations, we show that $\gamma(n) \le n$, and for relative perturbations we give a set of matrices $A \in M_n(R)$ with $\gamma(n) = n$. For a class of matrices including M-matrices we prove $\gamma(n) = 1$; i.e., $\sigma(A,E) = \rho(\,|A^{-1}| \cdot E)^{-1}$.

In section 6 the results are extended to obtain an explicit upper bound on $\gamma(n)$ for general A and E, and in section 7 those bounds are quantified into (0.4). We close with the conjecture that (0.4) is valid for $\gamma(n) = n$ for all A, E. If this is true, the set of matrices given in section 5 would imply that inequality (0.4) with $\gamma(n) = n$ is sharp.

1. Notation. In the following we give some notation from matrix theory; cf., for example, [7], [5]. $V_n(\mathbb{R})$ denotes the set of vectors with n real components, $M_{m,n}(\mathbb{R})$ the set of real m-by-n matrices, and $M_n(\mathbb{R}) = M_{n,n}(\mathbb{R})$. The components of a matrix $A \in M_n(\mathbb{R})$ are referred to by A_{ij} or $A_{i,j}$. For short notation, components of A^{-1} are referred to by A^{-1}_{ij}. $(\mathbf{1})$ denotes a vector with all components equal to 1 and $(\mathbf{1})_{nn} \in M_n(\mathbb{R})$ the matrix with all columns equal to $(\mathbf{1})$.

Q_{kn} denotes the set of strictly increasing sequences of k integers chosen from $\{1,\dots,n\}$. For $\omega \in Q_{kn}$, we denote $\omega = (\omega_1,\dots,\omega_k)$. For $C \in M_n(\mathbb{R})$, $\omega \in Q_{kn}$, $C[\omega] \in M_k(\mathbb{R})$ denotes the k-by-k submatrix of C lying in rows and columns ω. A sequence $\zeta = (i_1,\dots,i_k)$, $k \ge 1$ of mutually different integers $i_\nu \in \{1,\dots,n\}$ is called a *cycle*. We identify the cycles $(i_1,\dots,i_k)$ and $(i_p,\dots,i_k,i_1,\dots,i_{p-1})$, where $1 \le p \le k$. It is $|\zeta| := k$. A *full cycle* ζ on $\{1,\dots,n\}$ is a cycle ζ with $|\zeta| = n$.

For $C \in M_n(\mathbb{R})$ and a cycle $\zeta = (i_1,\dots,i_k)$ on $\{1,\dots,n\}$, we put

$$\Pi_\zeta(C) := C_{i_1 i_2} \times \cdots \times C_{i_{k-1} i_k} \cdot C_{i_k i_1},$$

the *cycle product* for ζ. Note the last factor in the product. Therefore, $|\Pi_\zeta(C)|^{1/|\zeta|}$ is the geometric mean of the elements of the cycle ζ. Each diagonal element C_{ii} is a cycle product of the cycle (i). (Here our definition differs from Engel and Schneider [4].)

With one exception, throughout the paper *absolute value* and *comparison* are used *componentwise*. For example, for $A, B \in M_n(\mathbb{R})$,

$$|A| \le B \quad \text{means} \quad |A_{ij}| \le B_{ij} \quad \text{for} \quad 1 \le i,j \le n.$$

The exception is cycles $\zeta = (i_1,\dots,i_k)$, where $|\zeta| = k$. The *singular values* of a matrix $A \in M_n(\mathbb{R})$ are denoted in decreasing order with increasing indices; i.e., $\sigma_1(A) \ge \cdots \ge \sigma_n(A) \ge 0$.

For $A, E \in M_n(\mathbb{R})$, $E \ge 0$, $\sigma(A,E)$ denotes the *componentwise distance*, weighted by E, *to the nearest singular matrix* (cf. (0.2)).

For finite $\sigma(A,E)$, the set of all matrices $\widetilde{A} \in M_n(\mathbb{R})$ with $|\widetilde{A} - A| \le \sigma(A,E) \cdot E$ is compact. For every nonsingular $\widetilde{A}$ there is a neighborhood of $\widetilde{A}$ consisting only of nonsingular matrices. Therefore,

$$\sigma(A,E) < \infty \;\Rightarrow\; \exists\, \delta A \in M_n(\mathbb{R}) : \;\; |\delta A| = \sigma(A,E) \cdot E \;\text{ and }\; A + \delta A \;\text{ singular},$$

showing that we are allowed to use a minimum in the definition (0.2) of $\sigma(A,E)$. ρ denotes the spectral radius, whereas ρ_0 denotes the *real spectral radius*:

$$B \in M_n(\mathbb{R}) : \;\; \rho_0(B) := \max\,\{\, |\lambda| \;\big|\; \lambda \in \mathbb{R} \;\text{ is an eigenvalue of }\; B \,\}.$$

If B has no real eigenvalues, we set $\rho_0(B) := 0$. I denotes the identity matrix of proper dimension; especially $I_k \in M_k(\mathbb{R})$ denotes the k-by-k identity matrix. A *signature matrix* S is a diagonal matrix with diagonal entries $+1$ or -1; i.e., $|S| = I$.

We frequently use standard results from matrix and Perron–Frobenius theory, such as

$$\begin{array}{l} A \in M_{nk}(\mathbb{R}), B \in M_{kn}(\mathbb{R}) \ \Rightarrow \\ \text{The set of nonzero eigenvalues of } AB \text{ and } BA \text{ are identical} \end{array} \tag{1.1}$$

(cf. Theorem 1.3.20 in [5]) and

$$\begin{array}{c} A \in M_n(\mathbb{R}) \text{ and } A \geq 0,\ x \in V_n(\mathbb{R}) \text{ with } x > 0 \ \Rightarrow \\ \min\limits_i \ \dfrac{(Ax)_i}{x_i} \leq \rho(A) \leq \max\limits_i \ \dfrac{(Ax)_i}{x_i}. \end{array} \tag{1.2}$$

The latter can be found in [2].

2. Finite perturbations for a linear system. Calculating bounds on $\sigma(A, E)$ can be motivated, for example, by looking at linear systems with finite perturbations of the input data. For a linear system $Ax = b$ consider the perturbed system $\widetilde{A}\widetilde{x} = \widetilde{b}$ with $\delta A := \widetilde{A} - A$, $\delta b := \widetilde{b} - b$, $\delta x := \widetilde{x} - x$. Then for nonsingular A,

$$A \cdot (I + A^{-1} \cdot \delta A) \cdot (\widetilde{x} - x) = \widetilde{A} \cdot (\widetilde{x} - x) = \widetilde{b} - \widetilde{A}x = \delta b - \delta A \cdot x. \tag{2.1}$$

If $\rho(A^{-1} \cdot \delta A) < 1$, then $I + A^{-1} \cdot \delta A$ and $\widetilde{A} = A \cdot (I + A^{-1} \cdot \delta A)$ are nonsingular, and (2.1) implies

$$\delta x = (I + A^{-1} \cdot \delta A)^{-1} \cdot A^{-1} \cdot (\delta b - \delta A \cdot x). \tag{2.2}$$

If $\rho\big(|A^{-1}| \cdot \Delta A\big) < 1$ then $I - |A^{-1}| \cdot \Delta A$ is an M-matrix. If the perturbations δA, δb are componentwise bounded by $|\delta A| \leq \Delta A$, $|\delta b| \leq \Delta b$, then (2.2) implies

$$|\delta x| \leq (I - |A^{-1}| \cdot \Delta A)^{-1} \cdot |A^{-1}| \cdot (\Delta b + \Delta A \cdot |x|\,). \tag{2.3}$$

For a given weight matrix ΔA, consider the set of matrices with componentwise distance from A weighted by ΔA not greater than σ:

$$\widetilde{A} \in U_\sigma(A, \Delta A) \ \Leftrightarrow \ |\widetilde{A} - A| \leq \sigma \cdot \Delta A.$$

For $\sigma \leq \rho(\,|A^{-1}| \cdot \Delta A)$, Perron–Frobenius theory yields

$$\rho(I - A^{-1} \cdot \widetilde{A}) = \rho\big(A^{-1} \cdot (A - \widetilde{A})\big) \leq \rho(\,|A^{-1}| \cdot \Delta A) < 1,$$

and therefore regularity of all $\widetilde{A} \in U_\sigma(A, \Delta A)$. The bound (2.3) requires $|A^{-1}| \cdot \Delta A$ to be convergent, whereas (2.2) is valid for $\rho(A^{-1} \cdot \delta A) < 1$. Therefore, we may ask how far a matrix $\widetilde{A}$ with $\rho(\,|A^{-1}| \cdot |\widetilde{A} - A|\,) \geq 1$ can be from the nearest singular matrix. An answer to this question shows how strong the assumption $\rho(\,|A^{-1}| \cdot \Delta A) < 1$ is.

3. Lower bounds on $\sigma(A, E)$. A simple and well-known lower bound on $\sigma(A, E)$ is

$$\frac{1}{\rho(\,|A^{-1}| \cdot E)} \leq \sigma(A, E) \quad \text{for all nonsingular} \quad A \in M_n(\mathbb{R}),\ 0 \leq E \in M_n(\mathbb{R}). \tag{3.1}$$

This can be seen using Perron–Frobenius theory and

$$\begin{aligned}\rho(\,|A^{-1}|\cdot E) < 1 &\Rightarrow \rho(A^{-1}\cdot\delta A) < 1 \quad \text{for all} \quad |\delta A| \le E \\ &\Rightarrow A+\delta A = A\cdot(I + A^{-1}\cdot\delta A) \quad \text{is nonsingular.}\end{aligned}$$

Another lower bound is (cf. [13, Theorem 1.8, p. 75])

$$\frac{\sigma_n(A)}{\sigma_1(E)} \le \sigma(A,E). \tag{3.2}$$

This can be generalized in the following way.

THEOREM 3.1. *Let $\|\cdot\|$ be a matrix norm subordinate to an absolute vector norm $\|\cdot\|$. Then for nonsingular $A \in M_n(\mathbb{R})$ and $0 \le E \in M_n(\mathbb{R})$,*

$$\frac{1}{\|A^{-1}\|\cdot\|E\|} \le \sigma(A,E). \tag{3.3}$$

Inequality (3.3) *is especially valid for all p-norms. For absolute norms such as the 1-norm and ∞-norm,*

$$\frac{1}{\|A^{-1}\|\cdot\|E\|} \le \frac{1}{\rho(\,|A^{-1}|\cdot E)} \le \sigma(A,E), \tag{3.4}$$

whereas for the 2-norm

$$\frac{1}{\|A^{-1}\|_2\cdot\|E\|_2} \le \frac{\sqrt{n}}{\rho(\,|A^{-1}|\cdot E)}. \tag{3.5}$$

Proof. To prove (3.3), let $\delta A \in M_n(\mathbb{R})$ with $|\delta A| \le \alpha\cdot E$ for $\alpha < (\,\|A^{-1}\|\cdot\|E\|\,)^{-1}$. The vector norm is absolute, implying $\|x\| = \|\,|x|\,\|$ and $|x| \le |y| \Rightarrow \|x\| \le \|y\|$ for $x, y \in V_n(\mathbb{R})$ (cf. [14, Theorem II.1.2]). Let $x \in V_n(\mathbb{R})$ with $\|x\| = 1$ and $\|\delta A\| = \|\delta A\cdot x\|$. Then

$$\begin{aligned}\|\delta A\| = \|\delta A\cdot x\| = \|\,|\delta A\cdot x|\,\| &\le \|\alpha\cdot E\cdot|x|\,\| \le \|\alpha\cdot E\|\cdot\|\,|x|\,\| \\ &= \alpha\cdot\|E\| < \|A^{-1}\|^{-1}.\end{aligned} \tag{3.6}$$

For every $0 \ne y \in V_n(\mathbb{R})$, $\|y\| \le \|A^{-1}\|\cdot\|Ay\|$ holds, and (3.6) yields

$$\|\delta A\cdot y\| \le \|\delta A\|\cdot\|y\| < \|A^{-1}\|^{-1}\cdot\|y\| \le \|Ay\|, \quad \text{and therefore} \quad (A+\delta A)\cdot y \ne 0.$$

Hence, $A+\delta A$ is nonsingular for $|\delta A| \le \alpha\cdot E$, and $\alpha < (\,\|A^{-1}\|\cdot\|E\|\,)^{-1}$, proving (3.3). For absolute matrix norms,

$$\rho(\,|A^{-1}|\cdot E) \le \|\,|A^{-1}|\cdot E\| \le \|A^{-1}\|\cdot\|E\|,$$

proving (3.4). For the 2-norm the following holds:

$$\rho(\,|A^{-1}|\cdot E) \le \|\,|A^{-1}|\,\|_2\cdot\|E\|_2 \le \|\,|A^{-1}|\,\|_F\cdot\|E\|_2 = \|A^{-1}\|_F\cdot\|E\|_2 \le \sqrt{n}\cdot\|A^{-1}\|_2\cdot\|E\|_2,$$

proving (3.5) and the theorem. ∎

Equation (3.4) shows that for absolute matrix norms such as the 1-norm or ∞-norm, the bound (3.3) cannot be better than (3.1). The 2-norm is not absolute, and (3.5) shows that the lower bound (3.2) for $\sigma(A,E)$ may be better up to a factor $\sqrt{n}$

than (3.1). In fact, we can identify a class of matrices for which this improvement is approximately achieved.

Let $Q \in M_n(\mathbb{R})$ be orthogonal, and consider absolute perturbations $E = (\mathbf{1})_{nn}$. Then (3.2) yields

$$\frac{\sigma_n(Q)}{\sigma_1\big((\mathbf{1})_{nn}\big)} = \frac{1}{n} \leq \sigma(Q,E).$$

On the other hand, $E = (\mathbf{1})_{nn} \in M_n(\mathbb{R})$ and $x = (\mathbf{1}) \in V_n(\mathbb{R})$ imply

$$\big\{\, |Q^{-1}| \cdot E \big\}^T \cdot x = E \cdot |Q| \cdot x = \Big(\sum_{i,j} |Q_{ij}| \Big) \cdot x,$$

and (1.2) yields $\rho(\, |Q^{-1}| \cdot E) = \sum_{i,j} |Q_{ij}|$. If Q is an orthogonalized random matrix with components uniformly distributed in $[-1,1]$, then $|Q_{ij}| \approx n^{-1/2}$. Thus, for the ratio between the two lower bounds (3.2) and (3.1) we obtain

$$\big\{\, \sigma_n(Q)/\sigma_1(E) \,\big\} \,/\, \big\{\, 1/\rho(\, |Q^{-1}| \cdot E) \,\big\} \approx n^{-1} \cdot n^2 \cdot n^{-1/2} = \sqrt{n}.$$

The same heuristic holds for $E = |Q|$; cf. [13]. For every Hadamard matrix ($H \in M_n(\mathbb{R})$ with $H^T H = n \cdot I$) the ratio is equal to $\sqrt{n}$. This sheds light on a possible quantity $\gamma(n)$ such that (0.4) holds. In section 5 we will prove $\gamma(n) \geq n$.

Example 3.2. The lower bound (3.2) may be arbitrarily weak. Consider

$$A = \begin{pmatrix} 2\varepsilon & -\varepsilon \\ -\varepsilon & 1 \end{pmatrix} \quad \text{and} \quad E = |A| \quad \text{for some } \varepsilon > 0.$$

A is a diagonally dominant M-matrix. As we will see in (5.5), A being an M-matrix implies equality in (3.1); i.e., $\sigma(A, |A|\,) = \rho(\, |A^{-1}| \cdot |A|\,) = 1 + 0(\sqrt{\varepsilon})$. On the other hand, $\sigma_2(A)/\sigma_1(\, |A|\,) = 2\varepsilon + 0(\varepsilon^2)$ underestimates $\sigma(A, |A|\,)$ arbitrarily. This corresponds to $\sigma_2(A) = 2\varepsilon + 0(\varepsilon^2)$. That means that the normwise distance in the 2-norm or Frobenius norm to the nearest singular matrix can be arbitrarily small compared to a componentwise distance.

4. A perturbation theorem for determinants. A lower bound on $\sigma(A,E)$ is obtained by proving the *regularity* of a set of matrices. This was done in section 3 by using spectral properties. To obtain an *upper* bound on $\sigma(A,E)$, we may construct a specific perturbation δA with $|\delta A| \leq \sigma_0 \cdot E$, $\sigma_0 \in \mathbb{R}$ such that $A + \delta A$ is singular. This proves $\sigma(A,E) \leq \sigma_0$. Another possibility for obtaining an upper bound on $\sigma(A,E)$ is the following. If $|\delta A| \leq \sigma_0 \cdot E$ and $\det(A) \cdot \det(A + \delta A) \leq 0$, then a continuity argument yields $\sigma(A,E) \leq \sigma_0$. Therefore, we state the following explicit formula for the relative change of the determinant of a matrix subject to a rank-k perturbation. It is a Sherman–Morrison–Woodbury-like perturbation formula for determinants.

LEMMA 4.1. *Let $A \in M_n(\mathbb{R})$ and $U, V \in M_{n,k}(\mathbb{R})$ be given. Then for nonsingular A,*

$$\det(A + UV^T) = \det(A) \cdot \det(I_k + V^T A^{-1} U), \tag{4.1}$$

where I_k denotes the k-by-k identity matrix.

Proof. It is

$$\det(A + UV^T) = \det(A) \cdot \det(I_n + A^{-1} U V^T).$$

Denoting the eigenvalues of $X \in M_n(\mathbb{R})$ by $\lambda_i(X)$ implies

$$\det(I_n + A^{-1}UV^T) = \prod_{i=1}^{n} \lambda_i(I_n + A^{-1}UV^T) = \prod_{i=1}^{n} \{1 + \lambda_i(A^{-1}UV^T)\}.$$

The set of nonzero eigenvalues of $A^{-1}UV^T$ and $V^TA^{-1}U$ are identical (see (1.1)), thus proving the lemma. □

This lemma has a nice corollary, which is interesting in itself.

COROLLARY 4.2. *Let $A \in M_n(\mathbb{R})$ and $u, v \in V_n(\mathbb{R})$. Then for nonsingular A,*

$$\det(A + uv^T) = \det(A) \cdot (1 + v^TA^{-1}u). \tag{4.2}$$

For arbitrary $A \in M_n(\mathbb{R})$ the following holds ($\mathrm{adj}(A)$ *denotes the adjoint of A):*

$$\det(A + uv^T) = \det(A) + v^T \cdot \mathrm{adj}(A) \cdot u. \tag{4.3}$$

The corollary shows that the relative change of the determinant is linear for rank-one perturbations of the matrix. The second well-known formula follows by a continuity argument using $A \cdot \mathrm{adj}(A) = \det(A) \cdot I$.

5. Upper bounds on $\sigma(A, E)$. The perturbation lemma for determinants given in section 4 allows for other lower bounds on $\sigma(A, E)$. The first result can be found in [9, Corollary 5.1, (iii)].

THEOREM 5.1. *Let $A \in M_n(\mathbb{R})$ be nonsingular and $E \in M_n(\mathbb{R})$ with $E \geq 0$. Then*

$$\sigma(A, E) \leq \frac{1}{\max_i \{\, |A^{-1}| \cdot E \,\}_{ii}}, \tag{5.1}$$

where 0^{-1} is interpreted as ∞.

Proof. Set $\alpha := \max_i \{\, |A^{-1}| \cdot E \,\}_{ii} \neq 0$ and let i be an index for which this maximum is achieved. Denote the $i\nu$th component of A^{-1} by $A^{-1}_{i\nu}$ and define $u \in V_n(\mathbb{R})$ by $u_\nu := -\alpha^{-1} \cdot \mathrm{sign}(A^{-1}_{i\nu}) \cdot E_{\nu i}$. Then

$$e_i^T \cdot A^{-1} \cdot u = -\alpha^{-1} \cdot \sum_{\nu=1}^{n} |A^{-1}_{i\nu}| \cdot E_{\nu i} = -1, \tag{5.2}$$

and Corollary 4.2 implies $\det(A + u \cdot e_i^T) = 0$. Now $|ue_i^T| \leq \alpha^{-1} \cdot E$ yields $\sigma(A, E) \leq \alpha^{-1}$. □

Example 5.2. The upper bound (5.1) can be arbitrarily weak. Consider

$$A = \begin{pmatrix} \varepsilon & 0 & 1 & 1 \\ 0 & \varepsilon & 1 & 1 \\ 1 & 1 & \varepsilon & 0 \\ 1 & 1 & 0 & \varepsilon \end{pmatrix}, E = |A| \text{ with } |A^{-1}| \cdot |A| \approx \begin{pmatrix} 1 & 1 & 1/\varepsilon & 1/\varepsilon \\ 1 & 1 & 1/\varepsilon & 1/\varepsilon \\ 1/\varepsilon & 1/\varepsilon & 1 & 1 \\ 1/\varepsilon & 1/\varepsilon & 1 & 1 \end{pmatrix}, \tag{5.3}$$

where the components of $|A^{-1}| \cdot |A|$ are accurate up to a relative error ε. Then (5.1) gives $\sigma(A, |A|\,) \leq 1 + 0(\varepsilon)$. On the other hand,

$$\det(A + \varepsilon \cdot \delta A) = 0 \quad \text{for} \quad \delta A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix},$$

showing that $\sigma(A, |A|) \le \varepsilon$.

In Theorem 5.1, a rank-one perturbation was used to prove (5.1). In a normwise sense, the minimum distance to the nearest singular matrix is achieved by a rank-one perturbation. This is no longer true for componentwise distances, as will be shown by the following example.

Example 5.3. According to Corollary 4.2, the smallest α such that $A + \alpha\widetilde{E}$ is singular with $|\widetilde{E}| \le \alpha \cdot |A|$ and $\mathrm{rank}(\widetilde{E}) = 1$ is given by $\alpha = |\widehat{\varphi}|^{-1}$, where $\widehat{\varphi}$ is an optimal value of the constraint optimization problem

$$\varphi(u, v) := v^T A^{-1} u = \text{Min!} \quad \text{subject to} \quad |uv^T| \le |A|.$$

In Example 5.2, partition the vectors $u, v \in V_4(\mathbb{R})$ into two vectors $U_i, V_i \in V_2(\mathbb{R})$, $i \in \{1, 2\}$, either having 2 components. That means $u = (U_1, U_2)^T$, $v = (V_1, V_2)^T$. Let

$$|uv^T| \le |A|; \quad \text{i.e.,} \quad U_i V_i^T \le \varepsilon \cdot I \quad \text{and} \quad U_i V_j^T \le (\mathbf{1})_{22} \quad \text{for} \quad 1 \le i,\ j \le 2,\ i \ne j.$$

The large elements of A^{-1} are in the upper left and lower right 2-by-2 blocks:

$$A^{-1} \approx \begin{pmatrix} X & Y \\ Y & X \end{pmatrix} \quad \text{with} \quad X = \frac{1}{2\varepsilon} \cdot \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \quad \text{and} \quad Y = \frac{1}{4} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

up to a relative error of the order ε. Therefore,

$$\begin{aligned} |v^T A^{-1} u| &\le |V_1^T X U_1| + |V_2^T X U_2| + |V_1^T Y U_2| + |V_2^T Y U_1| \\ &\le 2\varepsilon \cdot \sum |X_{ij}| + 2 \cdot \sum |Y_{ij}| \le 6. \end{aligned}$$

Thus, Corollary 4.2 implies that the minimum distance to the nearest singular matrix subject to rank-one perturbations weighted by $|A|$ is at least $1/6$ compared to $\sigma(A, |A|) \le \varepsilon$. This observation sheds light on the difficulties in calculating $\sigma(A, E)$ or finding upper bounds for it.

One may define the rank-k componentwise distance to the nearest singular matrix as follows:

$$\sigma_k(A, E) := \min\{\, \alpha \in \mathbb{R} \mid A + \widetilde{E} \text{ singular for } |\widetilde{E}| \le \alpha \cdot E \text{ and } \mathrm{rank}(\widetilde{E}) \le k \,\}.$$

We use $\mathrm{rank}(\widetilde{E}) \le k$ because E may be rank deficient. We have just seen in Example 5.3 that $\sigma_2(A, E)/\sigma_1(A, E)$ may be arbitrarily small.

Given the lower bound (3.1), one may ask whether there exist finite constants $\gamma(n) \in \mathbb{R}$ depending only on n such that

$$\frac{1}{\rho(\,|A^{-1}| \cdot E)} \le \sigma(A, E) \le \frac{\gamma(n)}{\rho(\,|A^{-1}| \cdot E)} \tag{5.4}$$

for all nonsingular $A \in M_n(\mathbb{R})$ and $0 \le E \in M_n(\mathbb{R})$. This question was raised in [3] and answered for some classes of matrices. The main purpose of this paper is to derive bounds for $\gamma(n)$. This will be done by using Lemma 4.1. For this purpose we need the following result by Rohn (for notation see section 1).

THEOREM 5.4. *For nonsingular $A \in M_n(\mathbb{R})$ and $0 \le E \in M_n(\mathbb{R})$ the following holds:*

$$\frac{1}{\max\limits_{S_1, S_2} \rho_0(S_1 A^{-1} S_2 E)} = \sigma(A, E),$$

where ρ_0 denotes the real spectral radius and the maximum is taken over all signature matrices. $1/0$ *is interpreted as* ∞.

Proof. See [10]. ∎

We start with a theorem bounding $\gamma(n)$ for general weight matrices E and identify a class of matrices with $\gamma(n) = 1$.

THEOREM 5.5. *For nonsingular* $A \in M_n(\mathbb{R})$ *and* $0 \le E \in M_n(\mathbb{R})$, *the following is true.*

(i) *Assume a matrix* $S \in M_n(\mathbb{R})$ *of rank one exists with*

$$S_{ij} = \begin{cases} +1 & \text{if} \quad A^{-1}_{ij} > 0, \\ -1 & \text{if} \quad A^{-1}_{ij} < 0, \\ +1 \text{ or } -1 & \text{if} \quad A^{-1}_{ij} = 0. \end{cases}$$

Then (5.4) *holds with* $\gamma(n) = 1$.

(ii) *If* $0 < \eta \le |E_{ij}| \le \zeta$ *for all* $1 \le i, j \le n$, *then* (5.4) *holds with* $\gamma(n) = n \cdot \zeta/\eta$.

Proof. Let $S = uv^T$ with $u, v \in V_n(\mathbb{R})$, $|u| = |v| = (\mathbf{1})$. Defining $S_1 = \mathrm{diag}(u)$, $S_2 = \mathrm{diag}(v)$, we have $S_1 A^{-1} S_2 = |A^{-1}|$, and Rohn's characterization in Theorem 5.4 proves the first part. Without loss of generality (w.l.o.g.) assume $\sigma(A, E) < \infty$. It is $\eta \cdot \|A^{-1}\|_\infty \le \max_i (\,|A^{-1}| \cdot E)_{ii}$ and $\rho(\,|A^{-1}| \cdot E) \le \|A^{-1}\|_\infty \cdot \|E\|_\infty \le n \cdot \zeta \cdot \|A^{-1}\|_\infty$. Thus, Theorem 5.1 proves the second part and therefore the theorem. ∎

For important classes of matrices such as inverse nonnegative matrices, among them all M-matrices, we already have a precise formula for $\sigma(A, E)$:

$$A \in M_n(\mathbb{R}) \text{ inverse nonnegative, } 0 \le E \in M_n(\mathbb{R}) \Rightarrow \ \sigma(A, E) = \frac{1}{\rho(\,|A^{-1}| \cdot E)}. \tag{5.5}$$

Example 5.6. If constants $\gamma(n)$ with (5.4) exist at all, we can give a lower bound on $\gamma(n)$ by means of the following. Define $A \in M_n(\mathbb{R})$ by

$$A := \begin{pmatrix} 1 & & & & & & s \\ 1 & 1 & & & 0 & & \\ & 1 & 1 & & & & \\ & & 1 & 1 & & & \\ & & & \ddots & \ddots & & \\ & 0 & & & & 1 & \\ & & & & & 1 & 1 \end{pmatrix} \quad \text{with} \quad s := (-1)^{n+1}. \tag{5.6}$$

The determinant of A calculates to

$$\det(A) = \prod_{i=1}^{n} A_{ii} + (-1)^{n+1} \cdot \Pi_\zeta(A) = 2, \quad \text{where } \zeta = (1, \ldots, n)$$

and $\prod_\zeta(A) = A_{12} \cdot A_{23} \times \cdots \times A_{n-1,n} \cdot A_{n1}$. If the elements of A are afflicted with relative perturbations, i.e., $E = |A|$, then only the 1's and s change. Therefore, any $\widetilde{A}$ with $|\widetilde{A} - A| \le \sigma \cdot |A|$ with $\sigma < 1$ is nonsingular, and therefore $\sigma(A, |A|\,) = 1$. On the other hand, $|A^{-1}| \cdot |A| = (\mathbf{1})_{nn}$ and $\rho(\,|A^{-1}| \cdot |A|\,) = n$. This proves the following lemma.

LEMMA 5.7. *If constants* $\gamma(n) \in \mathbb{R}$ *with* (5.4) *for every nonsingular* $A \in M_n(\mathbb{R})$ *and* $0 \le E \in M_n(\mathbb{R})$ *exist at all, then* $\gamma(n) \ge n$.

Next we show that $\gamma(n) \le n$ if E is of rank one. For the proof we use Corollary 4.2, which is a consequence of Lemma 4.1 for $k = 1$. In the remaining part of the

paper, we will extend this proof to $k > 1$ to obtain upper bounds for $\gamma(n)$ and for general A, E.

THEOREM 5.8. *Let nonsingular $A \in M_n(\mathbb{R})$ and $0 \le E \in M_n(\mathbb{R})$ with $E = uv^T$ for some $u, v \in V_n(\mathbb{R})$, $u, v \ge 0$ be given. Then*

$$\frac{1}{\rho(\,|A^{-1}| \cdot E)} \le \sigma(A, E) \le \frac{n}{\rho(\,|A^{-1}| \cdot E)} \ .$$

Proof. According to Theorem 5.4 and using (1.1),

$$\sigma(A, E)^{-1} = \max_{S_1, S_2} \rho_0(S_1 A^{-1} S_2 u v^T) = \max_{S_1, S_2} v^T S_1 A^{-1} S_2 u, \tag{5.7}$$

where the maximum is taken over all signature matrices S_1, S_2. For any i, $1 \le i \le n$, we can choose appropriate signature matrices S_1, S_2 such that $v^T S_1 A^{-1} S_2 u \ge v_i \cdot (\,|A^{-1}| \cdot u)_i$. Using (5.7) yields

$$\sigma(A, E)^{-1} \ge \max_i \ v_i \cdot (\,|A^{-1}| \cdot u)_i.$$

On the other hand, using (1.1),

$$\rho(\,|A^{-1}| \cdot E) = \rho(\,|A^{-1}| \cdot uv^T) = v^T \cdot |A^{-1}| \cdot u \le n \cdot \max_i \ v_i \cdot (\,|A^{-1}| \cdot u)_i. \qquad \square$$

COROLLARY 5.9. *For nonsingular $A \in M_n(\mathbb{R})$ and absolute perturbations, i.e., $E = (\mathbf{1})_{nn}$, estimation (5.4) holds with $\gamma(n) = n$.*

For absolute perturbations $E = (\mathbf{1})_{nn} = ee^T$ where $e := (\mathbf{1}) \in V_n(\mathbb{R})$, (5.7) implies $\sigma(A, E)^{-1} \ge \|A^{-1}\|_p$ for $p \in \{1, \infty\}$. Because $\|A^{-1}\|_2^2 \le \|A^{-1}\| \cdot \|A^{-1}\|_\infty$ this holds also for $p = 2$. Using $\|E\|_p = n$ for $p \in \{1, 2, \infty\}$ and the left inequality of Theorem 5.8, we have the following corollary.

COROLLARY 5.10. *For nonsingular $A \in M_n(\mathbb{R})$, $p \in \{1, 2, \infty\}$ and $E := (\mathbf{1})_{nn}$, we have*

$$\frac{1}{\|A^{-1}\|_p \cdot \|E\|_p} \le \sigma(A, E) \le \frac{n}{\|A^{-1}\|_p \cdot \|E\|_p}.$$

6. Estimation of $\gamma(n)$. To make further progress in the estimation of $\gamma(n)$ we show that for nonsingular A, $\sigma(A, E)$ depends continuously on A and E. Using this, we can restrict the class of matrices A and E to matrices with only nonzero components. For the proof we cannot use a simple continuity argument on $\rho_0(S_1 A^{-1} S_2 E)$ in connection with Theorem 5.4. This is because the search domain is restricted by E and the (in absolute value) largest *real* eigenvalue may be multiple and become complex under arbitrarily small perturbations.

LEMMA 6.1. *For nonsingular $A \in M_n(\mathbb{R})$, $\sigma(A, E)$ depends continuously on A and E.*

Proof. For $\sigma(A, E) = \infty$ we show that $\sigma(\widetilde{A}, \widetilde{E})$ becomes unbounded for $\widetilde{A} \to A$, $\widetilde{E} \to E$. A compactness and continuity argument shows that for every finite $0 < c \in \mathbb{R}$,

$$\forall\ |e| \le c \cdot E: \quad |\det(A + e)| \ge \delta > 0.$$

For every $\widetilde{A}$, $\widetilde{E}$ close enough to A, E, this implies $|\det(\widetilde{A} + \widetilde{e})| \ge \delta/2 > 0$ for every $|\widetilde{e}| \le c \cdot \widetilde{E}$, and hence $\sigma(\widetilde{A}, \widetilde{E}) > c$.

Assume $\sigma := \sigma(A,E) < \infty$. We will show that for small enough $\varepsilon > 0$, there exists some $\delta > 0$ such that both of the following statements are true:

$$\forall\ e \in M_n(\mathbb{R}):\quad |e| \le (\sigma-\varepsilon)\cdot E \quad\Rightarrow\quad \det(A)\cdot\det(A+e) > \delta, \tag{6.1}$$

$$\exists\ e \in M_n(\mathbb{R}):\quad |e| \le (\sigma+\varepsilon)\cdot E \ \text{ and } \ \det(A)\cdot\det(A+e) < -\delta. \tag{6.2}$$

Equation (6.1) is seen as follows. For $\varepsilon > 0$, the set of matrices $A+e$ with $|e| \le (\sigma-\varepsilon)\cdot E$ is nonempty and compact. Hence, $\det(A)\cdot\det(A+e)$ achieves a minimum on this set. By the definition of σ, this minimum is positive. To see (6.2), observe that $\det(A)\cdot\det(A+e) \ge 0$ for all $|e| \le \sigma\cdot E$. For any index pair i,j, the determinant $\det(A+\varepsilon\cdot e_i e_j^T)$ depends linearly on ε. Now proceed as follows. There is some e such that $A+e$ is singular and $|e| = E$. If for an index pair i,j the determinant $\det(A+e)$ is independent on e_{ij}, then replace e_{ij} by 0. At each step of this process, $\det(A+e) = 0$ and $|e| \le E$. The definition of $\sigma(A,E) < \infty$ and $\det(A) \ne 0$ imply that during this process we must arrive at some e and an index pair k, l, such that $\det(A+e)$ is not constant when changing e_{kl}. Then defining $e' \in M_n(\mathbb{R})$ by $e'_{ij} := e_{ij}$ for $(i,j) \ne (k,l)$ and $e'_{kl} := e_{kl}\cdot(1+\varepsilon')$ for small $\varepsilon' > 0$ proves (6.2).

Now the continuity of the determinant implies for $\widetilde{A}$, $\widetilde{E}$ close enough to A, E,

$$\forall\ |\widetilde{e}| \le (\sigma-\varepsilon)\cdot\widetilde{E}:\quad \det(\widetilde{A})\cdot\det(\widetilde{A}+\widetilde{e}) > \delta/2 \quad\text{and}$$

$$\exists\ |\widetilde{e}| \le (\sigma+\varepsilon)\cdot\widetilde{E}:\quad \det(\widetilde{A})\cdot\det(\widetilde{A}+\widetilde{e}) < -\delta/2,$$

and therefore $\sigma(A,E) - \varepsilon < \sigma(\widetilde{A},\widetilde{E}) < \sigma(A,E) + \varepsilon$. □

COROLLARY 6.2. *If* (5.4) *holds for each* $E > 0$*, then it holds for each* $E \ge 0$.

Our goal in this section is to prove the following upper bound for $\sigma(A,E)$. The quantities φ_t occurring in this estimation will be quantified and estimated in section 7.

PROPOSITION 6.3. *Let* $A, E \in M_n(\mathbb{R})$ *with* A *nonsingular and* $E \ge 0$ *be given. Define recursively* $\varphi_1 := 1$, $\varphi_2 := 1$, *and* $\varphi_t \in \mathbb{R}$, $2 < t \in \mathbb{N}$ *to be the (unique) positive root of*

$$P_t(x) \in \mathbb{R}[x] \quad\text{with}\quad P_t(x) := x^{t-1} - x^{t-2} - \sum_{\nu=1}^{t-1} \varphi_\nu^\nu \cdot x^{t-1-\nu}. \tag{6.3}$$

Then

$$\sigma(A,E) \le \frac{n\cdot\varphi_n}{\rho(\,|A^{-1}|\cdot E)}. \tag{6.4}$$

Therefore, the quantities $\gamma(n)$ *defined in* (5.4) *satisfy*

$$\gamma(1) = 1, \quad \gamma(2) = 2, \quad \text{and}$$
$$n \le \gamma(n) \le n\cdot\varphi_n. \tag{6.5}$$

The proof is divided into several parts and needs a preparatory lemma. First, we will construct a specific rank-k perturbation to be able to apply Lemma 4.1 to bound $\gamma(n)$ for general A, E. We use the same principle as in the proof of Theorem 5.1 adapted to rank-k perturbations.

LEMMA 6.4. *Let nonsingular $A \in M_n(\mathbb{R})$ and $0 \le E \in M_n(\mathbb{R})$ be given, and set $C := |A^{-1}| \cdot E$. For $1 \le k \le n$ define*

$$i' := \begin{cases} i+1 & \text{for } 1 \le i < k, \\ 1 & \text{for } i = k, \end{cases} \tag{6.6}$$

and $U, V \in M_{n,k}(\mathbb{R})$ by

$$U_{\nu i'} := \operatorname{sign}(A^{-1}_{i\nu}) \cdot E_{\nu i'} \quad \text{and} \quad V_{\mu i} := \delta_{\mu i}$$

for $1 \le \mu, \nu \le n$, $1 \le i \le k$, and the Kronecker symbol δ. Set $\widetilde{C} := V^T A^{-1} U$. Then

(i) $|\widetilde{C}| \le C[\omega]$ *for* $\omega = (1, \ldots, k)$,
(ii) $\widetilde{C}_{ii'} = C_{ii'}$ *for* $1 \le i \le k$,
(iii) $|UV^T| \le E$,
(iv) $\sigma(A, E) \le \{\rho_0(\widetilde{C})\}^{-1}$, *where* 0^{-1} *is interpreted as* ∞.

Proof. For $1 \le i, j \le k$ it follows that

$$|\,(V^T A^{-1} U)_{ij}| = \left| \sum_{\nu=1}^{n} \sum_{\mu=1}^{n} V_{\mu i} A^{-1}_{\mu\nu} U_{\nu j} \right| \le \sum_{\nu=1}^{n} |A^{-1}_{i\nu}| \cdot E_{\nu j} = C_{ij},$$

and therefore $|\widetilde{C}| \le C[\omega]$ and (i). For $1 \le i \le k$ the following holds:

$$\widetilde{C}_{ii'} = (V^T A^{-1} U)_{ii'} = \sum_{\nu=1}^{n} \sum_{\mu=1}^{n} V_{\mu i} A^{-1}_{\mu\nu} U_{\nu i'} = \sum_{\nu=1}^{n} A^{-1}_{i\nu} \cdot \operatorname{sign}(A^{-1}_{i\nu}) \cdot E_{\nu i'} = C_{ii'},$$

and therefore (ii). For $1 \le \mu, \nu \le n$ the following holds:

$$|\,(UV^T)_{\nu\mu}| = \left| \sum_{i=1}^{k} U_{\nu i} V_{\mu i} \right|,$$

such that $|(UV^T)_{\nu\mu}| = E_{\nu\mu}$ for $1 \le \mu \le k$, and $|(UV^T)_{\nu\mu}| = 0$ for $k+1 \le \mu \le n$. This proves (iii). For $\lambda := \rho_0(\widetilde{C}) > 0$, it is $\det(\lambda \cdot I - s \cdot \widetilde{C}) = 0$ for $s = -1$ or $s = 1$. Lemma 4.1 implies

$$\det(A - s \cdot \lambda^{-1} \cdot UV^T) = \det(A) \cdot \det(I_k - s \cdot \lambda^{-1} \cdot V^T A^{-1} U) = 0.$$

Together with (iii) and the definition (0.2) of $\sigma(A, E)$, this proves (iv) and the theorem. ∎

Our aim is to construct a rank-k perturbation of A with large real spectral radius. Then Lemma 4.1 allows us to give an upper bound on $\sigma(A, E)$. A first step is the following first generalization of Theorem 5.1. It will later yield the precise value for $\gamma(2)$.

THEOREM 6.5. *Let $A \in M_n(\mathbb{R})$ be nonsingular and $E \in M_n(\mathbb{R})$ with $E \ge 0$. For $C := |A^{-1}| \cdot E$ it holds that*

$$\sigma(A, E) \le \frac{1}{\max\limits_{i,j} \sqrt{C_{ij} \cdot C_{ji}}}. \tag{6.7}$$

Proof. For $i = j$, (6.7) was proven in Theorem 5.1. Reordering of indices puts the cycle (i, j), for which the maximum in (6.7) is achieved, into the cycle (1,2),

and Lemma 6.4 proves for $i \neq j$ the existence of a 2-by-2 matrix $\widetilde{C} = \left(\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right)$ with $0 \leq \beta = C_{ij}$, $0 \leq \gamma = C_{ji}$, $|\alpha| \leq C_{ii}$, $|\delta| \leq C_{jj}$, and $\sigma(A,E) \leq \rho_0(\widetilde{C})^{-1}$. If $|\alpha\delta| \geq \beta\gamma$, then $\sqrt{C_{ii} \cdot C_{jj}} \geq \sqrt{C_{ij} \cdot C_{ji}}$ and Theorem 5.1 yields (6.7). Otherwise, $\det(\widetilde{C}) < 0$. The characteristic polynomial of $\widetilde{C}$ is $\lambda^2 - \operatorname{trace}(\widetilde{C}) \cdot \lambda + \det(\widetilde{C})$, so that the eigenvalues of $\widetilde{C}$ are $\frac{1}{2} \cdot \{ \operatorname{trace}(\widetilde{C}) \pm \sqrt{\operatorname{trace}(\widetilde{C})^2 - 4 \cdot \det(\widetilde{C})} \} = \frac{1}{2} \cdot \{\alpha + \delta \pm \sqrt{(\alpha - \delta)^2 + 4\beta\gamma}\}$ and are both real. The absolute value of one of them is not less than $\sqrt{\beta\gamma}$; i.e., $\sigma(A,E) \leq \rho_0(\widetilde{C})^{-1} \leq (\beta\gamma)^{-1/2}$. ∎

The idea of the proof of Theorem 6.5 is the following: for a given cycle of C of length 2, a suitable rank-two perturbation of A is constructed which allows us to prove an upper bound of $\sigma(A,E)$ by using Lemma 6.4. In the following we will carry this idea to cycles of C of length k, $1 \leq k \leq n$.

First, we will identify a class of matrices for which we can give explicit *lower* bounds for their real spectral radius. The class of matrices is constructed in such a way that the matrices given in Lemma 6.4 can be used to bound $\sigma(A,E)$ from above.

LEMMA 6.6. *Let nonnegative $C \in M_k(\mathbb{R})$ and some $0 < a \in \mathbb{R}$ be given. Define $\varphi_1 := 1$, $\varphi_2 := 1$, and for $t > 2$ define recursively $\varphi_t \in \mathbb{R}$ to be the positive zero of*

$$P_t(x) \in \mathbb{R}[x] \quad \text{with} \quad P_t(x) := x^{t-1} - x^{t-2} - \sum_{\nu=1}^{t-1} \varphi_\nu^\nu \cdot x^{t-1-\nu}. \tag{6.8}$$

Suppose

$$\forall\ 1 \leq \mu < k \quad \forall\ \overline{\omega} \in \Gamma_{\mu k} : \quad |\Pi_{\overline{\omega}}(C)|^{1/\mu} \leq \varphi_\mu \cdot a, \tag{6.9}$$

and for $\omega = (1, \ldots, k)$,

$$|\Pi_\omega(C)|^{1/k} \geq \varphi_k \cdot a. \tag{6.10}$$

Then, for i' defined as in (6.6) *and every $\widetilde{C} \in M_k(\mathbb{R})$ with*

$$|\widetilde{C}| \leq C \quad \text{and} \quad \widetilde{C}_{ii'} = C_{ii'} \quad \text{for} \quad 1 \leq i \leq k, \tag{6.11}$$

the following holds:

$$\rho_0(\widetilde{C}) \geq a.$$

Proof. The proof is divided into the following parts. First, we transform C into a standard form such that all $C_{ii'}$ in the cycle $(1, \ldots, k)$ in (6.10) are equal. Second, we bound C by a circulant and show regularity of that matrix and $\det(\widetilde{C} - \lambda I) \neq 0$ for all $0 \leq \lambda < a$. Finally, the sign of the determinant of any $\widetilde{C}$ with (6.11) is determined, from which the lemma follows.

The case $k = 1$ is trivial; for $k = 2$ the proof of $\rho_0(\widetilde{C}) \geq a$ is included in the proof of Theorem 6.5.

Assume $k > 2$, and set $b := |\Pi_\omega(C)|^{1/k}$. Direct computation shows that any similarity transformation of C by a diagonal matrix D leaves all cycle products invariant.

Thus (6.9) and (6.10) remain valid when replacing C by $D^{-1}CD$ for any diagonal D with positive diagonal entries. Define diagonal $D \in M_k(\mathbb{R})$ by

$$f_i := b^{-1} \cdot C_{ii'} \quad \text{and} \quad D_{ii} := \prod_{\nu=i}^{k} f_\nu \quad \text{for} \quad 1 \leq i \leq k.$$

We show that w.l.o.g. C can be replaced by $D^{-1}CD$. We have $f_i > 0$, and (6.10) implies $D_{11} = 1$. It is

$$(D^{-1}CD)_{ii'} = \left(\prod_{\nu=i}^{k} f_\nu^{-1}\right) \cdot C_{ii'} \cdot \left(\prod_{\nu=i'}^{k} f_\nu\right) = C_{ii'} \cdot f_i^{-1} = b. \tag{6.12}$$

If $\widetilde{C} \in M_k(\mathbb{R})$ is any matrix satisfying (6.11), then $|D^{-1}\widetilde{C}D| \le D^{-1}CD$, and (6.12) yields $(D^{-1}\widetilde{C}D)_{ii'} = (D^{-1}CD)_{ii'} = b$. Since the set of eigenvalues of $\widetilde{C}$ and $D^{-1}\widetilde{C}D$ is identical, we can restrict our attention to matrices $C \in M_n(\mathbb{R})$, $C \ge 0$ and

$$C_{ii'} = b \quad \text{for} \quad 1 \le i \le k. \tag{6.13}$$

Set

$$C = \begin{pmatrix} c_{1,1} & b & c_{1,k-1} & \dots & & c_{1,2} \\ c_{2,2} & c_{2,1} & b & c_{2,k-1} & \dots & c_{2,3} \\ & \dots & c_{3,1} & b & \dots & \\ c_{k-1,k-1} & c_{k-1,k-2} & & \dots & & b \\ b & c_{0,k-1} & & \dots & & c_{0,1} \end{pmatrix}. \tag{6.14}$$

Let $\mu \in \mathbb{N}$, $1 \le \mu < k$ be given and define $\overline{\omega} \in \Gamma_{\mu k}$ by $\overline{\omega} = (1, \dots, \mu)$. Then setting $q := a/b$, (6.9) implies

$$c_{\mu\mu} \cdot \prod_{i=1}^{\mu-1} b \le (\varphi_\mu \cdot a)^\mu \quad \text{and therefore} \quad c_{\mu\mu} \le b \cdot \varphi_\mu^\mu \cdot q^\mu.$$

Applying the same argument successively for $\overline{\omega} = \big(i, (i+1) \bmod \mu, \dots, (i+\mu) \bmod \mu\big)$ yields

$$c_{i,\mu} \le b \cdot \varphi_\mu^\mu \cdot q^\mu \quad \text{for all} \quad 0 \le i < k,\ 1 \le \mu < k. \tag{6.15}$$

Therefore,

$$C \le b \cdot \begin{pmatrix} c_1 & 1 & c_{k-1} & & & c_2 \\ c_2 & c_1 & 1 & c_{k-1} & \dots & c_3 \\ & & c_1 & 1 & & \\ & \dots & & \dots & \dots & \\ c_{k-1} & c_{k-2} & & \dots & & 1 \\ 1 & c_{k-1} & & & & c_1 \end{pmatrix} =: b \cdot \overline{C} \tag{6.16}$$

with $c_\mu := \varphi_\mu^\mu \cdot q^\mu$ for $1 \le \mu < n$.

Let $\widetilde{C} \in M_k(\mathbb{R})$ with (6.11) be given, and let $\lambda \in \mathbb{R}$ with $0 \le \lambda < a$. Next we show that all matrices $\widetilde{C} - \lambda I$ are nonsingular. By assumption (6.11) and using (6.16),

$$|\widetilde{C} - \lambda I| \le C + \lambda \cdot I \le b \cdot \overline{C} + \lambda I \quad \text{and} \quad (\widetilde{C} - \lambda I)_{ii'} = \widetilde{C}_{ii'} = C_{ii'} = b. \tag{6.17}$$

By (6.16) and (6.8), using $q := a/b \le \varphi_k^{-1}$ from (6.10) and $\varphi_2 = 1$, we have for $k \ge 3$,

$$\begin{aligned} \lambda + b \cdot \sum_{\nu=1}^{k-1} c_\nu &< b \cdot \left\{ q + \sum_{\nu=1}^{k-1} \varphi_\nu^\nu \cdot q^\nu \right\} \le b \cdot \left\{ \varphi_k^{-1} + \sum_{\nu=1}^{k-1} \varphi_\nu^\nu \cdot \varphi_k^{-\nu} \right\} \\ &= b \cdot \varphi_k^{-k+1} \cdot \left\{ \varphi_k^{k-2} + \sum_{\nu=1}^{k-1} \varphi_\nu^\nu \cdot \varphi_k^{k-\nu-1} \right\} = b \cdot \varphi_k^{-k+1} \cdot \varphi_k^{k-1} = b. \end{aligned} \tag{6.18}$$

This shows that the element $b = \widetilde{C}_{ii'} = C_{ii'}$ strictly dominates the sum of the absolute values of the other components in each row of $C + \lambda I$ and of $\widetilde{C} - \lambda I$. That means that multiplication by a suitable permutation matrix produces a strictly diagonally dominant matrix and proves regularity of every $\widetilde{C} - \lambda I$ with $\widetilde{C}$ satisfying (6.11) and $0 \le \lambda < a$.

We proved that for every $\widetilde{C} \in M_k(\mathbb{R})$ with (6.11) the determinant of $\widetilde{C} - \lambda I$ is nonzero for $0 \le \lambda < a$. Therefore, the value of the characteristic polynomial $p(\lambda) = \det(\lambda I - \widetilde{C})$ of $\widetilde{C}$ has the same sign for $0 \le \lambda < a$. Now $p(\lambda) \to +\infty$ for $\lambda \to +\infty$. Therefore, the lemma is proven if we can show $p(0) < 0$, because in this case the characteristic polynomial must intersect with the real axis for some $\lambda^* \ge a$, thus proving $\rho_0(\widetilde{C}) \ge \lambda^* \ge a$.

We already proved that *every* matrix $\widetilde{C}$ satisfying (6.11) is nonsingular. Therefore $\operatorname{sign}\big(p(0)\big) = \operatorname{sign}\big(\det(-B)\big)$ for every matrix B with $|B| \le C$ and $B_{ii'} = C_{ii'} = b$. Define

$$B_{ij} := \begin{cases} C_{ii'} & \text{for } j = i', \\ 0 & \textit{otherwise.} \end{cases}$$

Then $\operatorname{sign}\big(\det(B)\big) = (-1)^{k+1}$, and therefore $\operatorname{sign}\big(p(0)\big) = (-1)^{2k+1} = -1$. The theorem is proven. ∎

Example 6.7. One can show that, at least for odd n, the bounds in Lemma 6.6 are sharp in the sense that there are examples with equality in (6.9) and (6.10) such that $\widetilde{C}$ with (6.11) exists with $\rho_0(\widetilde{C}) = a$. Consider

$$C := \begin{pmatrix} a & b & c \\ c & a & b \\ b & c & a \end{pmatrix} \text{ with } b := \varphi_3 \cdot a \text{ and } c := a/\varphi_3 \text{ and } \widetilde{C} := \begin{pmatrix} -a & b & -c \\ -c & -a & b \\ b & -c & -a \end{pmatrix}.$$

Then $C_{11} = \varphi_1 \cdot a = a$, $\sqrt{C_{12}C_{21}} = \varphi_2 \cdot a = a$, and $(C_{12}C_{23}C_{31})^{1/3} = \varphi_3 \cdot a$. $\widetilde{C}$ is a circulant, and its eigenvalues compute to $P(\varepsilon^k)$, $k = 0, 1, 2$, where $\varepsilon = e^{2\pi i/3}$ and $P(x) = bx^2 - cx - a$ (cf. [7]). It is $b - c - a = b(1 - \varphi_2^2 q^2 - \varphi_1 q) = b \cdot \varphi_3^{-1} = a$ with $q := a/b$. The other two eigenvalues are complex; thus $\rho_0(\widetilde{C}) = a$. The example extends to odd $n \in \mathbb{N}$.

The combination of Lemma 6.4, Theorem 6.5, and Lemma 6.6 gives the key to constructing a rank-k perturbation of A to achieve an upper bound for $\sigma(A, E)$. The following theorem is the generalization of Theorems 5.1 and 6.5 for cycles of length k, $1 \le k \le n$.

THEOREM 6.8. *Let $A, E \in M_n(\mathbb{R})$ with nonsingular A and $E \ge 0$ be given and define $C := |A^{-1}| \cdot E$. For $1 \le k \le n$ and any $\omega \in \Gamma_{kn}$ set*

$$0 \ne \tau := \big(\Pi_\omega(C)\big)^{1/k}. \tag{6.19}$$

Then for φ_k defined as in Lemma 6.6*,*

$$\sigma(A, E) \le \varphi_k/\tau.$$

In other words, φ_k divided by the geometric mean of the elements of any cycle of C bounds $\sigma(A, E)$ from above.

Proof. Let some $\omega \in \Gamma_{kn}$ and τ from (6.19) be given and set $a := \tau/\varphi_k$. If for $k = 1$ or $k = 2$ there exists some $\overline{\omega} \in \Gamma_{kn}$ with $\{\Pi_{\overline{\omega}}(C)\}^{1/k} \ge \varphi_k \cdot a$, then $\varphi_1 = \varphi_2 = 1$ and Theorems 5.1 and 6.5 imply $\sigma(A, E) \le a^{-1} = \varphi_k/\tau$. Therefore, we may assume

$\{\Pi_{\overline{\omega}}(C)\}^{1/k} < \varphi_k \cdot a$ for all $\overline{\omega} \in \Gamma_{kn}$ and $k \in \{1,2\}$. Hence, there is some $m \in \mathbb{N}$, $2 \le m \le k$ such that

$$\forall\ 1 \le \mu < m \quad \forall\ \overline{\omega} \in \Gamma_{\mu n}: \quad (\Pi_{\overline{\omega}}(C))^{1/\mu} \le \varphi_\mu \cdot a$$

and

$$\exists\ \widetilde{\omega} \in \Gamma_{mn}: \quad (\Pi_{\overline{\omega}}(C))^{1/m} \ge \varphi_m \cdot a.$$

After a suitable rearrangement of indices we may assume $\widetilde{\omega} = (1,\dots,m)$, and Lemma 6.4 yields a matrix $\widetilde{C} \in M_m(\mathbb{R})$ with properties (i) and (ii) of Lemma 6.4 and $\sigma(A,E) \le \rho_0(\widetilde{C})^{-1}$. But Lemma 6.6 shows that for all such matrices $\rho_0(\widetilde{C}) \ge a = \tau/\varphi_m$. Regarding $m \le k$, the theorem is proven if we can show

$$t \in \mathbb{N} \quad \Rightarrow \quad \varphi_t \le \varphi_{t+1}. \tag{6.20}$$

We know $\varphi_1 = \varphi_2 = 1$; from definition (6.8) we see that $\varphi_3 = 1 + \sqrt{2}$, and for $t \ge 3$

$$P_{t+1}(x) = x \cdot P_t(x) - \varphi_t^t.$$

Hence $P_{t+1}(\varphi_t) < 0$ and $\varphi_{t+1} > \varphi_t$. The theorem is proven. □

Theorem 6.8 reduces the problem of finding upper bounds of $\sigma(A,E)$ to one of finding proper cycles of some length k of $|A^{-1}| \cdot E$ with a large geometric mean corresponding to a suitable rank-k perturbation. This is done in the following proof of Proposition 6.3.

Proof of Proposition 6.3. Corollary 6.2 allows us to assume $E > 0$. Therefore, $|A^{-1}| \cdot E$ is positive, and Perron–Frobenius theory yields the existence of a positive eigenvector $x \in V_n(\mathbb{R})$ with $|A^{-1}| \cdot E \cdot x = \rho(\,|A^{-1}| \cdot E) \cdot x$, $\rho(\,|A^{-1}| \cdot E) > 0$. Define the diagonal matrix $D_x \in M_n(\mathbb{R})$ by $(D_x)_{ii} := x_i$. We may replace A by $A \cdot D_x$ and E by $E \cdot D_x$, because for any nonsingular diagonal matrices D_1, D_2, $\sigma(A,E) = \sigma(D_1 A D_2, D_1 E D_2)$. This is because $|\delta A| \le \sigma \cdot E$ *iff* $|D_1 \cdot \delta A \cdot D_2| \le \sigma \cdot |D_1 E D_2|$ and $A + \delta A$ is singular *iff* $D_1 A D_2 + D_1 \cdot \delta A \cdot D_2$ is singular (cf. [3]). Then

$$C := |\,(A \cdot D_x)^{-1}| \cdot E \cdot D_x = D_x^{-1} \cdot |A^{-1}| \cdot E \cdot D_x \quad \text{and} \quad C \cdot (\mathbf{1}) = \rho(\,|A^{-1}| \cdot E) \cdot (\mathbf{1}).$$

This means C is a multiple of a row stochastic matrix. Set $\rho := \rho(\,|A^{-1}| \cdot E)$.

Denote an index of the maximal component of C in row i by m_i. Then either $\{\, m_i \mid 1 \le i \le n \,\} = \{1,\dots,n\}$ or there is a cycle $m_j, m_{j+1}, \dots, m_{j+k-1}, m_{j+k} = m_j$ of length k. This means that with a suitable renumbering there is some $k \in \mathbb{N}$, $1 \le k \le n$ such that for the upper left k-by-k principal submatrix of C the following holds:

$$C_{ii'} \ge \rho/n \quad \text{for} \quad 1 \le i \le k, \tag{6.21}$$

where i' is defined as in (6.6).

Then Theorem 6.8, (6.20), and (6.21) imply for $\omega = (1,\dots,k)$ that

$$\sigma(A,E) \le \varphi_k \cdot \{\Pi_\omega(C)\}^{-1/k} \le n \cdot \varphi_k/\rho \le n \cdot \varphi_n/\rho. \qquad \square$$

In the remaining section 7, we will replace the bound (6.5) by giving explicit bounds for $\gamma(n)$ depending only on n. An asymptotic bound will be given as well.

7. Explicit bounds for $\gamma(n)$. The main result of section 6 is the upper bound (6.5) in Proposition 6.3. This bound is given in terms of φ_k, the positive zeros of the polynomial P_t defined in (6.3). In the remaining part of the paper we will give bounds on $\gamma(n)$ showing the dependence on n by a simple function. Moreover, the asymptotic behavior of $\gamma(n)$ for $n \to \infty$ is given.

The polynomials $P_t(x) \in \mathbb{R}[x]$ defined in (6.3) satisfy

$$P_t(x) = x^{t-1} - x^{t-2} - \sum_{\nu=1}^{t-1} \varphi_\nu^\nu \cdot x^{t-1-\nu} \quad \text{and} \quad P_t(\varphi_t) = 0 \quad \text{for} \quad t > 2. \tag{7.1}$$

Therefore, for $n \geq 3$,

$$\varphi_n^{-1} + \sum_{i=1}^{n-1} \varphi_i^i \cdot \varphi_n^{-i} = 1. \tag{7.2}$$

By (6.20), $x + \sum_{i=1}^{n-1} \varphi_i^i \cdot x^i$ is strictly increasing for $x > 0$. Hence, for $x > 0$,

$$x + \sum_{i=1}^{n-1} \varphi_i^i \cdot x^i \leq 1 \quad \text{implies} \quad x \leq \varphi_n^{-1}; \quad \text{that is,} \quad \varphi_n \leq x^{-1}. \tag{7.3}$$

We are aiming for a bound of the form

$$\varphi_k \leq c \cdot k^\alpha \tag{7.4}$$

for some constants c and α. To determine c and α, we notice that if (7.4) is satisfied for $1 \leq k < n$, then

$$\sum_{i=1}^{n-1} \left(\frac{i}{n}\right)^{\alpha_i} \leq 1 - c^{-1} \cdot n^{-\alpha} \quad \text{implies} \quad \varphi_n \leq c \cdot n^\alpha. \tag{7.5}$$

This is because the left-hand side of (7.5) yields

$$1 \geq c^{-1} \cdot n^{-\alpha} + \sum_{i=1}^{n-1} i^{\alpha i} \cdot n^{-\alpha i} \geq (c \cdot n^\alpha)^{-1} + \sum_{i=1}^{n-1} \varphi_i^i \cdot (c \cdot n^\alpha)^{-i}$$

and (7.3) implies $(c \cdot n^\alpha)^{-1} \leq \varphi_n^{-1}$.

Therefore, our first step is to derive upper bounds for

$$\sum_{i=1}^{n-1} \sigma_i \quad \text{with} \quad \sigma_i := \left(\frac{i}{n}\right)^{i\alpha}. \tag{7.6}$$

σ_i depends on n and α. We use the abbreviation σ_i for fixed n and α and omit extra parameters for better readability. In order to estimate the sum (7.6), we will split it into three parts which will be bounded individually. For $i \geq 1$ the following holds:

$$\frac{\sigma_{i+1}}{\sigma_i} = \left(\frac{i+1}{n}\right)^{(i+1)\alpha} \cdot \left(\frac{n}{i}\right)^{(i+1)\alpha} \cdot \left(\frac{i}{n}\right)^{\alpha} = \left(\frac{i}{n}\right)^{\alpha} \cdot \left(1 + \frac{1}{i}\right)^{(i+1)\alpha} > \left(\frac{i}{n}\right)^{\alpha} \cdot e^\alpha,$$

and therefore

$$\sigma_i < \left(\frac{n}{i \cdot e}\right)^{\alpha} \cdot \sigma_{i+1} \quad \text{for} \quad i \geq 1. \tag{7.7}$$

For all $\beta \in \mathbb{R}$ with $1 < \beta < e$ and $k := \lceil \frac{n\cdot\beta}{e} \rceil$ it holds that $k-1 < \frac{n\cdot\beta}{e} \le k$. Then (7.7) gives

$$\sum_{i=k}^{n-1} \sigma_i < \sigma_{n-1} + \sum_{i=k}^{n-2} \left(\frac{n}{i\cdot e}\right)^\alpha \cdot \sigma_{i+1} = \sigma_{n-1} + \sum_{i=k+1}^{n-1} \left(\frac{n}{(i-1)\cdot e}\right)^\alpha \cdot \sigma_i,$$

and therefore

$$\sigma_{n-1} - \sigma_k > \sum_{i=k+1}^{n-1} \left\{1 - \left(\frac{n}{(i-1)\cdot e}\right)^\alpha\right\} \cdot \sigma_i \ge \sum_{i=k+1}^{n-1} \left\{1 - \left(\frac{n}{k\cdot e}\right)^\alpha\right\} \cdot \sigma_i,$$

and $\frac{n}{k\cdot e} \le \beta^{-1}$ yields

$$\sigma_{n-1} - \sigma_k > (1-\beta^{-\alpha}) \cdot \sum_{i=k+1}^{n-1} \sigma_i.$$

Now, $\beta > 1$ and $\alpha \ge 0$ imply $(1-\beta^{-\alpha})^{-1} > 1$. Therefore,

$$\sum_{i=k}^{n-1} \sigma_i < (1-\beta^{-\alpha})^{-1} \cdot \sigma_{n-1} = (1-\beta^{-\alpha})^{-1} \cdot \left(\frac{n-1}{n}\right)^{(n-1)\alpha} =: \mu_n \tag{7.8}$$

holds for every $\alpha \ge 0$, $1 < \beta < e$, and $k := \lceil \frac{n\beta}{e} \rceil$. This is the first part of sum (7.6) for a suitable k to be determined. Define

$$f(x) := \left(\frac{x}{n}\right)^{x\alpha} \quad \text{with} \quad f'(x) = \left(\frac{x}{n}\right)^{x\alpha} \cdot \left\{\alpha \cdot \ln \frac{x}{n} + \alpha\right\}.$$

For $x > 0$, $f(x)$ has exactly one minimum at $x = \frac{n}{e}$. Then $f(i) = \sigma_i$ shows that

$$\sigma_k \ge \sigma_l \quad \text{for} \quad 1 \le k \le l \le \frac{n}{e} \quad \text{and} \quad \sigma_k \le \sigma_l \quad \text{for} \quad \frac{n}{e} \le k \le l \le n-1. \tag{7.9}$$

Set $M := \lceil n/e \rceil$. Then $k = \lceil \frac{n\beta}{e} \rceil$ satisfies $M \le k \le n$, and (7.9) implies for $n \ge 3$ that

$$\begin{aligned} \sum_{i=M}^{k-1} \sigma_i &\le (k-M)\cdot\sigma_{k-1} < \left(\frac{n\beta}{e} + 1 - \frac{n}{e}\right) \cdot f\left(\frac{n\beta}{e}\right) < \frac{n\beta}{e} \cdot f\left(\frac{n\beta}{e}\right) \\ &\le n \cdot \left(\frac{\beta}{e}\right)^{n\frac{\beta\alpha}{e}+1} =: \nu_n. \end{aligned} \tag{7.10}$$

This is the second part of sum (7.6). Finally, (7.9) implies that

$$\sum_{i=1}^{M-1} \sigma_i < \left(\frac{1}{n}\right)^\alpha + \left(\frac{2}{n}\right)^{2\alpha} + \left(\frac{3}{n}\right)^{3\alpha} + \left(\frac{4}{n}\right)^{4\alpha} \cdot \left(\frac{n}{e}\right) =: \xi_n, \tag{7.11}$$

which is the third part of sum (7.6). Inequalities (7.8), (7.10), and (7.11) together yield

$$n^{-\alpha} + \sum_{i=1}^{n-1} \left(\frac{i}{n}\right)^{i\alpha} \le n^{-\alpha} + \mu_n + \nu_n + \xi_n \quad \text{for} \quad n \ge 3. \tag{7.12}$$

Next, we show that all three sequences μ_n, ν_n, ξ_n are decreasing for large enough n. $\left(1+\frac{1}{n}\right)^n$ is monotonically increasing for $n \geq 1$; therefore, for $n \geq 2$,

$$\left(\frac{n+1}{n}\right)^{n\alpha} \geq \left(\frac{n}{n-1}\right)^{(n-1)\alpha} \;\Rightarrow\; \mu_{n+1} \leq \mu_n.$$

Suppose

$$n_0 \geq \left\{ \left(\frac{e}{\beta}\right)^{\frac{\beta\alpha}{e}} - 1 \right\}^{-1}. \tag{7.13}$$

Then for $n \geq n_0$,

$$1+\frac{1}{n} \leq \left(\frac{e}{\beta}\right)^{\frac{\beta\alpha}{e}} \;\Rightarrow\; (n+1)\cdot\left(\frac{\beta}{e}\right)^{\frac{\beta\alpha}{e}} \leq n \;\Rightarrow (n+1)\cdot\left(\frac{\beta}{e}\right)^{(n+1)\frac{\beta\alpha}{e}+1} \leq n\cdot\left(\frac{\beta}{\alpha}\right)^{n\frac{\beta\alpha}{e}+1},$$

and therefore $\nu_{n+1} \leq \nu_n$ for $n \geq n_0$ with n_0 satisfying (7.13). Finally, for $n \geq 1$ and $\alpha > 0.25$, $1-4\alpha < 0$ and therefore

$$(n+1)^{1-4\alpha} \leq n^{1-4\alpha} \;\Rightarrow\; \left(\frac{4}{n+1}\right)^{4\alpha}\cdot\left(\frac{n+1}{e}\right) \leq \left(\frac{4}{n}\right)^{4\alpha}\cdot\left(\frac{n}{e}\right) \;\Rightarrow\; \xi_{n+1} \leq \xi_n.$$

Summarizing, this proves the following lemma.

LEMMA 7.1. *Define $\varphi_1 := 1$, $\varphi_2 := 1$, and recursively φ_n to be the positive zero of $P_n(x)$ given in* (6.3). *Let constants $c, \alpha \in \mathbb{R}$, $\alpha \geq \ln 2$ and $3 \leq n_0 \in \mathbb{N}$ be given with $\varphi_n \leq c\cdot n^\alpha$ for $n < n_0$. If a constant $\beta \in \mathbb{R}$, $1 < \beta < e$ exists such that* (7.13) *is satisfied and μ_n, ν_n, ξ_n defined in* (7.8), (7.10), (7.11) *satisfy*

$$n^{-\alpha} + \mu_n + \nu_n + \xi_n \leq 1 \quad \text{for} \quad n = n_0, \tag{7.14}$$

then

$$\varphi_n \leq c\cdot n^\alpha \quad \text{for all} \quad n \in \mathbb{N}.$$

Proof. Equation (7.4) is satisfied for $1 \leq k < n$, and (7.12) and (7.14) prove the left-hand side of (7.5) for $n = n_0$, and therefore (7.4) for $k = n$. The quantities $n^{-\alpha}$, μ_n, ν_n, and ξ_n are decreasing for increasing n. Thus, (7.14) and therefore (7.4) are valid for all $n \geq n_0$. By assumption, $\varphi_n \leq c\cdot n^\alpha$ for $n < n_0$ as well. ∎

For example, for $\beta := 2.697$, $\alpha := 0.7$, and $n_0 := 3000$, one checks by explicit calculation that $\varphi_n \leq 2.321\cdot n^\alpha$ for $1 \leq n \leq n_0$. The lower bound (7.13) for n_0 is less than 183, $\mu_n < 0.992$, $\nu_n < 0.0003$, $\xi_n < 0.0038$, and $n^{-\alpha} < 0.0038$ for $n = n_0$. This proves the following result.

COROLLARY 7.2. *For all $n \geq 1$, $\varphi_n \leq 2.321\cdot n^{0.7}$. The difference $2.321\cdot n^{0.7} - \varphi_n$ is less than 2.8 for $1 \leq n < 20$ and less than 2.0 for $20 \leq n \leq 2000$.*

Summarizing, Corollary 7.2, Proposition 6.3, and Lemma 5.7 prove the following result.

PROPOSITION 7.3. *Let $A, E \in M_n(\mathbb{R})$ with nonsingular A and $E \geq 0$ be given. Then for all $n \geq 1$,*

$$\frac{1}{\rho(\,|A^{-1}|\cdot E)} \leq \sigma(A,E) \leq \frac{\gamma(n)}{\rho(\,|A^{-1}|\cdot E)},$$

with

$$n \le \gamma(n) \le 2.321 \cdot n^{1.7}.$$

The lower bound for $\gamma(n)$ is sharp.[1]

Finally, we will show the asymptotic behavior of upper bounds for $\gamma(n)$. Let $\alpha := \ln(2+2\eta)$, $\eta > 0$. For any $1 < \beta < e$ and $n \to \infty$,

$$n^{-\alpha} \to 0, \quad \mu_n \to (1-\beta^{-\alpha})^{-1} \cdot e^{-\alpha}, \quad \nu_n \to 0, \quad \text{and} \quad \xi_n \to 0.$$

For $\ln \beta := (2+\eta)/(2+2\eta)$, a short computation yields

$$(1-\beta^{-\alpha})^{-1} \cdot e^{-\alpha} = \frac{2+\eta}{2+4\eta+2\eta^2} < 1.$$

Hence, for this β and large enough n_0, (7.13) holds and

$$n^{-\alpha} + \mu_n + \nu_n + \xi_n < 1 \quad \text{for all} \quad n \ge n_0.$$

Therefore, for large enough c with $\varphi_n \le c \cdot n^\alpha$ for $n < n_0$, Lemma 7.1 implies that $\varphi_n \le c \cdot n^\alpha$ for all $n \in \mathbb{N}$. Using $\alpha > \ln 2$ proves the following.

PROPOSITION 7.4. *Let $\gamma(n)$ be defined as follows:*

$$\gamma(n) := \inf\{\, \sigma(A,E) \cdot \rho(|A^{-1}| \cdot E) \mid A \in M_n(\mathbb{R}) \text{ nonsingular and } 0 \le E \in M_n(\mathbb{R}) \,\}.$$

Then $\gamma(n)$ is finite for all $n \in \mathbb{N}$. Moreover, for any $\varepsilon > 0$ there exists some $n_0 \in \mathbb{N}$ such that for all $n \ge n_0$ the following holds:

$$n \le \gamma(n) \le n^{1+\ln 2+\varepsilon}. \tag{7.15}$$

The lower bound in (7.15) *is sharp.*

In his paper [3], Demmel showed that for the Bauer–Skeel condition number $\kappa(A,E) := \| \, |A^{-1}| \cdot E\|$ with any p-norm, $1 \le p \le \infty$, the following holds:

$$\frac{1}{\rho(\,|A^{-1}| \cdot E)} = \frac{1}{\min\limits_D \kappa(AD,\, ED)}\,,$$

where the minimum is taken over all diagonal D. Thus, Propositions 7.3 and 7.4 prove that the componentwise relative distance to the nearest singular matrix for any weight matrix $E \ge 0$ is not too far from the reciprocal of the smallest condition number achievable by column scaling. The evidence presented in this paper leads us to the following conjecture.

CONJECTURE 7.5. *For all nonsingular $A \in M_n(\mathbb{R})$ and $0 \le E \in M_n(\mathbb{R})$ the following holds:*

$$\frac{1}{\rho(\,|A^{-1}| \cdot E)} \le \sigma(A,E) \le \frac{n}{\rho(\,|A^{-1}| \cdot E)}\,. \tag{7.16}$$

If the conjecture is true, Lemma 5.7 *shows that both bounds are sharp (see also the note to Proposition* 7.3*).*

[1] In the meantime it has been shown by the author that $n \le \gamma(n) \le (3+2\sqrt{2}) \cdot n$. This is included in *Almost sharp bounds for the componentwise distance to the nearest singular matrix*, to appear in LAMA. It uses extensively the author's results in *Theorems of Perron–Frobenius type for matrices without sign restrictions*, to appear in LAA.

Acknowledgment. The author wishes to thank Jiri Rohn for many helpful comments.

REFERENCES

[1] F. Bauer, *Optimally scaled matrices*, Numer. Math., 5 (1963), pp. 73–87.
[2] L. Collatz, *Einschließungssatz für die charakteristischen Zahlen von Matrizen*, Math. Z., 48 (1942), pp. 221–226.
[3] J. Demmel, *The componentwise distance to the nearest singular matrix*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 10–19.
[4] G. Engel and H. Schneider, *The Hadamard-Fischer inequality for a class of matrices defined by eigenvalue monotonicity*, Linear and Multilinear Algebra, 4 (1976), pp. 155–176.
[5] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, London, 1985.
[6] W. Kahan, *Numerical linear algebra*, Canad. Math. Bull., 9 (1966), pp. 757–801.
[7] M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Dover, New York, 1992.
[8] S. Poljak and J. Rohn, *Checking robust nonsingularity is NP-hard*, Math. Control Signals Systems, 6 (1993), pp. 1–9.
[9] J. Rohn, *Nearness of Matrices to Singularity*, KAM Series on Discrete Mathematics and Combinatorics, 1988.
[10] J. Rohn, *Systems of linear interval equations*, Linear Algebra Appl., 126 (1989), pp. 39–78.
[11] J. Rohn, *Interval matrices: Singularity and real eigenvalues*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 82–91.
[12] S. Rump, *Estimation of the sensitivity of linear and nonlinear algebraic problems*, Linear Algebra Appl., 153 (1991), pp. 1–34.
[13] S. Rump, *Verification methods for dense and sparse systems of equations*, in Topics in Validated Computations—Studies in Computational Mathematics, J. Herzberger, ed., Elsevier, Amsterdam, The Netherlands, 1994, pp. 63–136.
[14] G. Stewart and J. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 104–118, January 1997

009

STABILITY ISSUES IN THE FACTORIZATION OF STRUCTURED MATRICES*

MICHAEL STEWART† AND PAUL VAN DOOREN‡

Abstract. This paper provides an error analysis of the generalized Schur algorithm of Kailath and Chun [*SIAM J. Matrix Anal. Appl.*, 15 (1994), pp. 114–128]—a class of algorithms which can be used to factorize Toeplitz-like matrices, including block-Toeplitz matrices, and matrices of the form T^TT, where T is Toeplitz. The conclusion drawn is that if this algorithm is implemented with hyperbolic transformations in the factored form which is well known to provide numerical stability in the context of Cholesky downdating, then the generalized Schur algorithm will be stable. If a more direct implementation of the hyperbolic transformations is used, then it will be unstable. In this respect, the algorithm is analogous to Cholesky downdating; the details of implementation of the hyperbolic transformations are essential for stability. An example which illustrates this instability is given. This result is in contrast to the ordinary Schur algorithm for which an analysis by Bojanczyk, Brent, De Hoog, and Sweet [*SIAM J. Matrix Anal. Appl.*, 16 (1995), pp. 40–57] shows that the stability of the algorithm is not dependent on the implementation of the hyperbolic transformations.

Key words. Schur algorithm, structured matrices, Toeplitz matrices, stability

AMS subject classifications. 65F05, 65F30, 15A06, 15A23

PII. S089547989528692X

1. Introduction.

The Schur algorithm is a popular and fast method for the Cholesky factorization of a square, positive-definite Toeplitz matrix T. It performs reliably, and in [3] it was shown to be stable in the sense that if the algorithm runs to completion and $\hat{C}$ is the computed Cholesky factor, $\|T - \hat{C}^T\hat{C}\|$ is guaranteed to be small. This paper will perform a similar stability analysis which applies to several special cases of the generalized Schur algorithm [10]. In its full generality, the generalized Schur algorithm can be adapted to the factorization of a wide variety of structured matrices. The analysis given here is primarily of interest for block-Toeplitz and Toeplitz-block matrices, as well as for matrices of the form T^TT, where T is rectangular and Toeplitz. The key notion behind the general algorithm is the concept of displacement rank [10].

One of the most significant examples is the Cholesky factorization of T^TT. This factor is also the factor R in the QR factorization of the rectangular Toeplitz matrix T, and the obvious application of this fact to the solution of Toeplitz least squares problems is explored in [1]. However, the analysis given there assumes the use of the algorithm presented in [2] rather than the generalized Schur algorithm. The basic idea is to obtain R without bothering about finding Q, thus avoiding any problems associated with the loss of orthogonality which are common to all fast Toeplitz QR algorithms. The method of seminormal equations, possibly with iterative refinement, can then be used to find the least squares solution. The resulting equations are $R^TRx = T^Tb$ and a weak stability result can be given concerning their solution provided that for the computed R, $\|R^TR - T^TT\|$ is kept small. As will be shown

* Received by the editors May 30, 1995; accepted for publication (in revised form) by N. J. Higham January 9, 1996. This work was supported by ARPA grant 60NANB2D1272 and NSF grant CCR-9209349.

http://www.siam.org/journals/simax/18-1/28692.html

† Coordinated Science Laboratory, University of Illinois, Urbana, IL 61801 (stewart@monk.csl.uiuc.edu).

‡ Department of Mathematical Engineering, Université Catholique de Louvain, Louvain-la-Neuve, Belgium (vandooren@anma.ucl.ac.be).

here, this can be done using the generalized Schur algorithm as well as the algorithm in [2]. Consequently, the observations concerning Toeplitz least squares problems in [1] can be adapted to an alternate approach—that of finding R from the generalized Schur algorithm and then solving the seminormal equations to obtain the solution.

A comparison of the analysis in this paper can also be made with the analysis of the Schur algorithm given in [3]. Despite some major similarities, the conclusions of this paper will be, in one very important respect, quite different: the implementation details of the Schur algorithm are less critical to stability than those of the generalized Schur algorithm. Both can be implemented using hyperbolic transformations of the form

$$H = \frac{1}{\sqrt{1-\rho^2}} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

with $|\rho| < 1$ and where the transformations are applied with straightforward matrix multiplication; however there are reasons to be concerned about the stability of this approach. In many algorithms, such as Cholesky downdating [4], this is not a good idea. A downdating algorithm based on such transformations will not be stable unless the transformations are implemented in factored form,

$$H = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{1-\rho^2}} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ 0 & 1 \end{bmatrix}.$$

It is worth noting that the manner in which the elements of these factors are computed is also important for stability. Details can be found in [4] and a correct implementation can be found in section 3. Such a seemingly minor difference in implementation makes the difference between stability and instability for Cholesky downdating, but the Schur algorithm is more robust—it is stable using either approach. Unfortunately, the generalized Schur algorithm will be shown to be analogous to the downdating problem in this regard; the factored transformations are essential for stability. A proof of the stability of the factored hyperbolic approach will be given in section 4 while an example which illustrates the instability of the other approach will be given in section 5. The stability proof in [1] also requires use of the factored hyperbolic transformations. This is not surprising since the presentation in [2] recasts the problem in the form of a downdating problem.

Finally, in section 5, some comments will be made about the nature of the instability. The propagation of errors is essentially stable regardless of which implementation is used; the difference is in the size of the local errors. From the fact that the error propagation is stable, it is possible to show that there will be many instances in which the instability does not completely destroy the accuracy of the computed Cholesky factors. This is true even for quite ill-conditioned problems. Numerical examples will be given to support this claim.

2. Displacement rank and stable computation of the initial generators. Given an $n \times n$ symmetric, positive-definite matrix A, which is not necessarily Toeplitz, we define the displacement of A to be

$$D_A = A - Z_A A Z_A^T, \tag{1}$$

where Z_A is strictly lower triangular (a matrix with zeros on the diagonal). The restriction of symmetry on A can easily be relaxed to allow A to be Hermitian, but for simplicity we deal only with the real symmetric case here.

In [10], the only additional restriction on Z is that it be a fast, $O(n)$ rather than $O(n^2)$, process to multiply a vector by Z. Otherwise, the factorization algorithm will have a complexity of $O(n^3)$ instead of $O(n^2)$. In addition to this, we will also make two assumptions which will guarantee the stability of the algorithm and simplify the analysis. For stability it will be necessary to assume that $\|Z\|_2 \leq 1$. Otherwise, repeated multiplication by Z will magnify errors. Further, to simplify the analysis, it will be assumed that a vector can be multiplied by Z without incurring any error. These two assumptions essentially limit Z to be a shift or a block shift. Since these are the forms that Z takes for Toeplitz least squares problems and block-Toeplitz Cholesky factorization, respectively, these assumptions are reasonable, even if they do remove a good deal of generality. The second assumption is not essential if the multiplication is done in a stable manner, but there don't seem to be any common examples to which the analysis could be made to apply by dropping it.

To be more specific about typical examples of Z, if A is square and Toeplitz or has the form T^TT for some rectangular Toeplitz matrix T, we choose $(Z_T)_{ij} = 1$ if $i = j + 1$ and $(Z_T)_{ij} = 0$ otherwise. A block-Toeplitz matrix B would require Z_B to be a block shift, with the ones on the subdiagonal replaced with identity matrices. The significance of the analysis here will be for algorithms based on displacements involving Z_T and Z_B.

It is well known that if T is a symmetric Toeplitz matrix, normalized to have ones on the diagonal, $(T)_{ij} = t_{i-j}$ with $t_i = t_{-i}$ and $t_0 = 1$, then D_T will be an indefinite rank 2 matrix and

$$D_T = T - Z_T T Z_T^T = G_T^T \Sigma_T G_T,$$

where

$$G_T = \begin{bmatrix} 1 & t_1 & \cdots & t_{n-1} \\ 0 & t_1 & \cdots & t_{n-1} \end{bmatrix} = \begin{bmatrix} u^T \\ v^T \end{bmatrix}$$

and $\Sigma_T = 1 \oplus -1$.

The vectors u and v are referred to as *generators* of T. Clearly, the matrix T is uniquely determined by these generators. Performing operations on the generators rather than on T directly is what allows efficient $O(n^2)$ algorithms for Toeplitz systems.

Assume that for some Z_A satisfying the previously specified restrictions A belongs to a class of positive-definite matrices for which D_A as defined by equation (1) has rank α. Clearly D_A will have a decomposition of the form

$$D_A = G^T \Sigma_A G,$$

where

$$G = \begin{bmatrix} u_{11} & u_{12}^T \\ u_{21} & U_{22} \\ v_{11} & v_{12}^T \\ v_{21} & V_{22} \end{bmatrix}$$

and $\Sigma_A = I_p \oplus I_q$. Here we take u_{11} and v_{11} to be scalars and u_{21} and v_{21} to be column vectors of size $p - 1$ and $q - 1$, respectively, with $p + q = \alpha$.

In section 3, we will give a summary of the general algorithm. More details and many special cases can be found in [10]. But first, since the numerical stability of

the overall factorization will depend on the stability of the process of finding the generators for A, we will briefly discuss how this can be done in a stable manner.

As already seen, the generators of a Toeplitz matrix can be found trivially with no error. For a symmetric block-Toeplitz matrix, the process is only slightly more complicated and the errors in the initial generators will not cause a problem with stability. The generators for a Toeplitz least squares problem can also be found in a reliable manner with a minimal amount of computation, and again the process is stable.

In general, the numbers p and q correspond, respectively, to the number of positive and negative eigenvalues of D. When this decomposition cannot be obtained trivially, as in the Toeplitz or block-Toeplitz case, it can be obtained through an eigenvalue decomposition or through Gaussian elimination with a symmetric pivoting strategy to obtain an LDL^T factorization of the displacement [7].

Of course, computing a full eigenvalue decomposition will slow down the overall algorithm. In the absence of any specific knowledge of the form which the generators will take, and in the case in which D can be computed with $O(n^2)$ complexity, computing α steps of the LDL^T decomposition will give a set of generators without destroying the $O(n^2)$ complexity of the algorithm. However, it is important to note that when truncating such a decomposition, the pivoting scheme which is chosen can be more critical to stability than when the decomposition is carried out completely. The fact that the Bunch–Parlett strategy is a backward stable method for computing LDL^T is well known [7]. The stability of the Bunch–Kaufman procedure has been established more recently in [9]. However, it is simple to verify numerically that the rank 2 matrix

$$\begin{bmatrix} \epsilon & \sqrt{\epsilon} & \epsilon \\ \sqrt{\epsilon} & \frac{\beta^2}{4} & \frac{\beta}{2} \\ \epsilon & \frac{\beta}{2} & \frac{\beta^2\epsilon - 4\beta\sqrt{\epsilon} + \beta^2}{\beta^2 - 4} \end{bmatrix},$$

where $\beta = (1+\sqrt{17})/8$ is a constant chosen to minimize element growth for symmetric pivoting and $\epsilon > 0$ is very small, leads to a large error if two steps of the Bunch–Kaufman approach are used to obtain a rank 2 LDL^T factorization. The Bunch–Parlett procedure will not have this problem. The reason for the difference can be found by looking at the sensitivity of the scalar Schur complement which is truncated by these two algorithms. An analysis of pivoted low rank Cholesky factorization which highlights the relevance of Schur complement sensitivity in the case of semidefinite matrices was given in [8]; the issues are the same for indefinite matrices, and the extension to the more complicated pivoting strategies is direct. The result can be summarized as follows: backward stability guarantees that if the computed Schur complement after two stages of the factorization process is small, then no large errors will be incurred in dropping it and terminating the factorization. Rank deficiency ensures that the exact value of the Schur complement will be zero, but if it is highly sensitive to perturbations in the original matrix, then the backward error can cause the computed value to become large. It is shown in [8] that the sensitivity of the Schur complement will depend on the size of the elements in L. A careful scrutiny of the two pivoting strategies shows that, although both are backward stable and give bounds on the norms of the Schur complements, only Bunch–Parlett gives bounds on the elements of L. However, this digression is included only for completeness to show that the process of obtaining the generators can always be done in a stable fashion. In practice, there is generally a simpler means of obtaining the generators—

for Toeplitz matrices, they can be obtained with no computation at all. A general factorization method would only be needed in the rare case in which D does not have a zero structure which makes the choice of generators obvious, or at least easy to compute. An example is the case of general non-Toeplitz matrices for which α equals two. The class of matrices for which $\alpha = 2$ is broader than the class of Toeplitz matrices, and D_A will not always have a zero structure which makes the process of finding the generators trivial.

3. The generalized Schur algorithm. Assuming that we have managed to find the generators for a structured matrix A, we can apply the generalized Schur algorithm to find a factorization $A = C^T C$. Much of the power and generality of this approach is due to the fact that it is a straightforward and intuitively clear generalization of the Schur algorithm; there is little essential difference between the cases $\alpha = 2$ and $\alpha > 2$. Since the former case is well known, the presentation here of the latter case can be kept brief. A more leisurely description can be found in [10].

Let $Z = Z_A$ and $Z_j = Z_A^j$. Since Z is strictly lower triangular, $Z_n = 0$, and from this it follows that

$$A = \sum_{j=0}^{n-1} Z_j (A - ZAZ^T) Z_j^T$$

or

$$A = \sum_{j=0}^{n-1} Z_j G^T \Sigma G Z_j^T. \tag{2}$$

For any transformation J such that $J^T \Sigma J = \Sigma$, we find that JG is also a set of generators for A. This follows immediately, since $G^T J^T \Sigma J G = G^T \Sigma G = D$.

From equation (2) and the positivity of A, we see that

$$0 < (A)_{1,1} = \left\| \begin{bmatrix} u_{11} \\ u_{12} \end{bmatrix} \right\|^2 - \left\| \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} \right\|^2.$$

This means that if a transformation J_1 of the form

$$J_1 = \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix},$$

where Q_1 and Q_2 are orthogonal, is applied to G to give

$$J_1 G = \begin{bmatrix} \hat{u}_{11} & \hat{u}_{12}^T \\ 0 & \hat{U}_{22} \\ \hat{v}_{11} & v_{12}^T \\ 0 & \hat{V}_{22} \end{bmatrix}$$

then $|\hat{u}_{11}| > |\hat{v}_{11}|$. Therefore it is possible to take $\rho = -\hat{v}_{11}/\hat{u}_{11}$ so that

$$J_2 J_1 G = \frac{1}{\sqrt{1-\rho^2}} \begin{bmatrix} 1 & 0 & \rho & 0 \\ 0 & I & 0 & 0 \\ \rho & 0 & 1 & 0 \\ 0 & 0 & 0 & I \end{bmatrix} J_1 G = \begin{bmatrix} \tilde{u}_{11} & \tilde{u}_{12}^T \\ 0 & \hat{U}_{22} \\ 0 & \tilde{v}_{12}^T \\ 0 & \hat{V}_{22} \end{bmatrix}.$$

Note that $J_1^T \Sigma J_1 = \Sigma$ and $J_2^T \Sigma J_2 = \Sigma$ so that $J_2 J_1 G$ is also a set of generators for A. We have proven that we can always find a set of generators for which the first column has only one nonzero element. Such generators are said to be proper.

This form of the generators is very useful. Equation (2) implies that the first row of A is not different from the first row of $G^T \Sigma G$. When the generators are in proper form, this means that

$$A(1,:) = \tilde{u}_{11} \begin{bmatrix} \tilde{u}_{11} & \tilde{u}_{12}^T \end{bmatrix}.$$

Clearly, the first row of the Cholesky factor of A will be

$$C(1,:) = \begin{bmatrix} \tilde{u}_{11} & \tilde{u}_{12}^T \end{bmatrix}.$$

If we let

$$A_S = A - \begin{bmatrix} \tilde{u}_{11} \\ \tilde{u}_{12} \end{bmatrix} \begin{bmatrix} \tilde{u}_{11} & \tilde{u}_{12}^T \end{bmatrix}$$

be the Schur complement of A, we see that

$$A_S - Z A_S Z^T = G^T \Sigma G - \begin{bmatrix} \tilde{u}_{11} \\ \tilde{u}_{12} \end{bmatrix} \begin{bmatrix} \tilde{u}_{11} & \tilde{u}_{12}^T \end{bmatrix} + Z \begin{bmatrix} \tilde{u}_{11} \\ \tilde{u}_{12} \end{bmatrix} \begin{bmatrix} \tilde{u}_{11} & \tilde{u}_{12}^T \end{bmatrix} Z^T.$$

This means that merely postmultiplying the first row of the generators

$$u^T = \begin{bmatrix} \tilde{u}_{11} & \tilde{u}_{12}^T \end{bmatrix}$$

by Z^T will give the generators of the Schur complement.

Additional unitary and hyperbolic transformations can now be used to introduce zeros into the second elements of these vectors to obtain the first row of the Cholesky factor of the Schur complement—the second row of the Cholesky factor of A. The process can be continued recursively to obtain the complete Cholesky factor of A. At each stage, the positivity of the Schur complements guarantees that an appropriate ρ can be found.

The resulting algorithm, with the hyperbolic transformations implemented in the stable form in MATLAB code, is as follows:

```
function [C]=gschur(G,n,p,q)
C=zeros(n,n); C(1,1:n)=G(1,1:n);
for i=1:n,
  for j=2:p
    Q=givens(G(1,i),G(j,i));
    G(1:j-1:j,i:n)=Q*G(1:j-1:j,i:n);
  end
  for j=2:q
    Q=givens(G(p+1,i),G(p+j,i));
    G(p+1:j-1:p+j,i:n)=Q*G(p+1:j-1:p+j,i:n);
  end
  s=G(p+1,i)/G(1,i);
  c=sqrt(G(1,i)^2-G(p+1,i)^2)/G(1,i);
  G(1,i+1:n)=([1 -s]*G(1:p:p+1,i+1:n)/c;
```

```
  G(p+1,i+1:n)=[-s c]*G(1:p:p+1,i+1:n);
  G(1,i)=sqrt(G(1,i)^2-G(p+1,i)^2);
  C(i,i:n)=G(1,i:n);
  G(1,i+1:n)=G(1,i:n-1);
end;
```

It was mentioned in section 1 that the computation of the elements in the factors is significant for stability. It is also important that the leading element of the first generator be computed separately, as shown here. The analysis in [4] does not apply to other implementations.

Although the algorithm given above will suffice to find the factors of a block-Toeplitz matrix, it is significantly different from some of the approaches given in the literature. The numerical properties are not different, but a more block-oriented perspective on block-Toeplitz factorization can be found in [6].

In the two-generator case, the matrix J_2 which zeros the first elements of all the generators except the first is almost uniquely defined by the constraints

$$H \begin{bmatrix} u_{11} \\ v_{11} \end{bmatrix} = \begin{bmatrix} x \\ 0 \end{bmatrix}$$

and $H^T \Sigma H = \Sigma$. The only possible variation is that each row of H can be multiplied by -1. This is not the case when there are more generators. The obvious example is the possibility of applying an arbitrary hyperbolic transformation acting on the last $\alpha - 1$ generators after the appropriate zeros have been introduced into their leading elements. This paper will show that when J is formed from a block diagonal orthogonal matrix and an appropriately implemented hyperbolic transformation, the algorithm will be stable.

4. Stability analysis. A stability analysis of the Schur algorithm for Toeplitz factorization is presented in [3]. The analysis can be broken into two parts: one which shows how local errors propagate through the algorithm and one which bounds the local errors. The propagation of errors is essentially the same for the generalized Schur algorithm, but the problem of bounding local errors is slightly more difficult. We will first modify results from [3] to apply to the general algorithm and then later develop new inequalities which will complete the analysis.

At the end of the kth stage of the algorithm, let the generators stored in the computer be

$$\tilde{G}_k = \begin{bmatrix} 0_{k-1}^T & \tilde{u}_{11,k} & \tilde{u}_{12,k}^T \\ 0_{p-1,k-1} & 0_{p-1} & \tilde{U}_{22,k} \\ 0_{k-1}^T & 0 & \tilde{v}_{12,k}^T \\ 0_{q-1,k-1} & 0_{q-1} & \tilde{V}_{22,k} \end{bmatrix} = \begin{bmatrix} \tilde{u}_{1,k}^T \\ \tilde{U}_{2,k} \\ \tilde{v}_{1,k}^T \\ \tilde{V}_{2,k} \end{bmatrix}.$$

Using MATLAB notation, let $\tilde{G}_{k,Z}$ be defined by

$$\tilde{G}_{k,Z}(1,:) = \tilde{G}_k(1,:)Z^T$$

and

$$\tilde{G}_{k,Z}(2:\alpha,:) = \tilde{G}_k(2:\alpha,:).$$

Also, let G_k and $G_{k,Z}$ be the generators which would result from exact computations.

The computed generators will satisfy

$$\tilde{G}_{k+1}^T \Sigma \tilde{G}_{k+1} = \tilde{G}_{k,Z}^T \Sigma \tilde{G}_{k,Z} + \epsilon F_k + O(\epsilon^2),$$

where ϵ is the machine precision and ϵF_k are errors incurred locally in computing the new form of the generators at step k through the use of orthogonal and hyperbolic transformations.

Since it is possible that G_1 will already be in proper form, we will assume that it is and treat any errors which appear in it separately from the other errors. Summing this equation from $k = 1$ to $k = n - 1$ and grouping the terms related to the top rows of $\tilde{G}_k$ and $\tilde{G}_{k,Z}$ in the left-hand side gives

$$\sum_{k=1}^{n-1} \left(\tilde{u}_{1,k+1} \tilde{u}_{1,k+1}^T - Z \tilde{u}_{1,k} \tilde{u}_{1,k}^T Z^T \right) = \sum_{k=1}^{n-1} \Big((\tilde{G}_{k,Z}(2:\alpha,:))^T \Sigma (\tilde{G}_{k,Z}(2:\alpha,:)) - (\tilde{G}_{k+1}(2:\alpha,:))^T \Sigma (\tilde{G}_{k+1}(2:\alpha,:)) + \epsilon F_k \Big) + O(\epsilon^2).$$

After simplifying the terms that cancel in the right-hand side, we obtain

$$\tilde{R}^T \tilde{R} - \tilde{u}_{1,1} \tilde{u}_{1,1}^T - Z^T \tilde{R}^T \tilde{R} Z + Z \tilde{u}_{1,n} \tilde{u}_{1,n}^T Z^T = (\tilde{G}_1(2:\alpha,:))^T \Sigma (\tilde{G}_1(2:\alpha,:)) - (\tilde{G}_n(2:\alpha,:))^T \Sigma (\tilde{G}_n(2:\alpha,:)) + \epsilon \sum_{k=1}^{n-1} F_k + O(\epsilon^2),$$

where

$$\tilde{R} = \begin{bmatrix} \tilde{u}_{1,1}^T \\ \tilde{u}_{1,2}^T \\ \vdots \\ \tilde{u}_{1,n}^T \end{bmatrix}.$$

Since $G_{n,Z} = 0$,

$$\tilde{R}^T \tilde{R} - Z^T R^T R Z = \tilde{u}_{1,1} \tilde{u}_{1,1}^T + (\tilde{G}_1(2:\alpha,:))^T \Sigma (\tilde{G}_1(2:\alpha,:)) + \epsilon \sum_{k=1}^{n-1} F_k + O(\epsilon^2).$$

Using this to find the displacement

$$(A - \tilde{R}^T \tilde{R}) - Z(A - \tilde{R}^T \tilde{R}) Z^T = G_1^T \Sigma G_1 - \tilde{G}_1^T \Sigma \tilde{G}_1 + \epsilon \sum_{k=1}^{n-1} F_k + O(\epsilon^2)$$

and applying equation (2) gives

$$A - \tilde{R}^T \tilde{R} = \sum_{j=0}^{n-1} Z_j [G_1^T \Sigma G_1 - \tilde{G}_1^T \Sigma \tilde{G}_1] Z_j^T - \epsilon \sum_{j=0}^{n-1} \sum_{k=1}^{n-j-1} Z_j F_k Z_j^T + O(\epsilon^2), \tag{3}$$

where the sum over k has been reduced by noting that $Z_j^T F_k Z_j^T = 0$ whenever $k > n - j - 1$. This shows that if we can bound errors in the computation of the initial generators and the local errors, then the algorithm will be stable. The errors in the initial generators are not a problem, since from D the generators can be obtained in

a backward stable manner using Bunch–Parlett pivoting to compute LDL^T or, more typically, in a more direct fashion. Since the methods for obtaining the generators may vary, in the analysis and error bounds which follow we will ignore this source of error and only concern ourselves with the effects of local errors due to the unitary and hyperbolic transformations.

The local errors are given by the expression

$$\epsilon F_k = \tilde{G}_{k+1}^T \Sigma \tilde{G}_{k+1} - \tilde{G}_{k,Z}^T \Sigma \tilde{G}_{k,Z} + O(\epsilon^2).$$

Because any bounds on the errors produced by the transformations will depend on the norm of the generators, it is essential to bound the generators.

THEOREM 4.1. *When the generators are computed by applying a sequence of plane rotations and a hyperbolic transformation, they satisfy*

$$\|G_k\|_F^2 \leq 2\sqrt{k-1}\|A\|_F + \|G_1\|_F^2.$$

Proof. Let $\hat{u}_1$ and $\hat{v}_1$ be the two generators u_1 and v_1 after orthogonal transformations have been performed on the generators from step $k-1$. Then

$$\begin{aligned}
\begin{bmatrix} u_{1,k}^T \\ v_{1,k}^T \end{bmatrix} &= \frac{1}{\sqrt{1-\rho^2}} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} \hat{u}_1^T \\ \hat{v}_1^T \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{1-\rho^2}} & \frac{\rho}{\sqrt{1-\rho^2}} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{u}_1^T \\ \hat{v}_1^T \end{bmatrix} \\
&= \begin{bmatrix} u_{1,k}^T \\ \rho u_{1,k}^T \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} \hat{u}_1^T \\ \hat{v}_1^T \end{bmatrix}.
\end{aligned}$$

Taking Frobenius norms gives

$$\begin{aligned}
\left\| \begin{bmatrix} u_{1,k}^T \\ v_{1,k}^T \end{bmatrix} \right\|_F^2 &= (1+\rho^2)\|u_{1,k}\|^2 + (1-\rho^2)\|\hat{v}_1\|^2 + 2\rho\sqrt{1-\rho^2} u_{1,k}^T \hat{v}_1 \\
&\leq \|u_{1,k}\|^2 + (|\rho|\|u_{1,k}\| + \sqrt{1-\rho^2}\|\hat{v}_1\|)^2 \\
&\leq 2\|u_{1,k}\|^2 + \|v_1\|^2.
\end{aligned}$$

To see why the final inequality is true, let $x = |\rho|$ and look for the minimum value of

$$f(x) = \|u_{1,k}\| + (x\|u_{1,k}\|^2 + \sqrt{1-x^2}\|\hat{v}_1\|)^2$$

on $0 \leq x \leq 1$. If

$$0 = f'(x) = 2\left(x\|u_{1,k}\| + \sqrt{1-x^2}\|\hat{v}_1\|\right)\left(\|u_{1,k}\| - \frac{x}{\sqrt{1-x^2}}\|\hat{v}_1\|\right),$$

then

$$x = \frac{\|u_{1,k}\|}{\sqrt{\|u_{1,k}\|^2 + \|\hat{v}_1\|^2}}$$

and

$$f(x) = 2\|u_{1,k}\|^2 + \|\hat{v}_1\|^2.$$

It is easy to see that f assumes lower values at $x = 0$ and $x = 1$, so this point is the only possible maximum. Since the plane rotations do not affect the norm of the generators and since $\|Z\|_2 \leq 1$,

$$\|G_k\|_F^2 \leq 2\|u_{1,k}\|^2 + \|G_{k-1}\|_F^2.$$

This inequality can be expanded recursively to give

$$\|G_k\|_F^2 \leq 2\sum_{j=2}^{n} \|u_{1,j}\|^2 + \|G_1\|_F^2 = 2\|C(2:k,:)\|_F^2 + \|G_1\|_F^2.$$

Finally, for an arbitrary positive semidefinite rank $k-1$ matrix with a factorization $A = C^T C$,

$$\|C\|_F^2 \leq \sqrt{k}\|A\|_F.$$

This follows from the fact that the Frobenius norm squared equals the sum of squares of the singular values and from a standard inequality relating the vector 2-norm and the vector 1-norm. This completes the proof of the theorem.

To complete a stability analysis all that is needed is to show that the orthogonal and hyperbolic transformations produce a local error, ϵF_k, which is proportional to the norm of the generators. Note that this does not necessarily refer to the norm of the generators prior to the transformation. In fact, in the case of the hyperbolic transformation, it is necessary to look at the norm of one of the generators which is produced by the hyperbolic transformation.

An error analysis of hyperbolic transformations is given in [4]. The result is that if the transformations are applied in factored form

$$\frac{1}{\sqrt{1-\rho^2}} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{1-\rho^2}} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ 0 & 1 \end{bmatrix} \tag{4}$$

then

$$\begin{bmatrix} \tilde{u}_{1,k+1}^T + \widehat{\Delta u}^T \\ \tilde{v}_{1,k+1}^T \end{bmatrix} = H \begin{bmatrix} \hat{u}_{1,k}^T \\ \hat{v}_{1,k}^T + \widehat{\Delta v}^T \end{bmatrix}. \tag{5}$$

The mixed error vectors $\widehat{\Delta u}$ and $\widehat{\Delta v}$ satisfy

$$\left\| \begin{bmatrix} \widehat{\Delta u}^T \\ \widehat{\Delta v}^T \end{bmatrix} \right\|_F \leq 6.25\epsilon \left\| \begin{bmatrix} \tilde{u}_{1,k+1}^T \\ \hat{v}_{1,k}^T \end{bmatrix} \right\|_F, \tag{6}$$

where ϵ is the unit roundoff.

In addition to this, there is a result in [11] concerning the application of plane rotations. In particular, there will exist orthogonal $\hat{Q}_1$ and $\hat{Q}_2$ such that

$$\left[\begin{array}{c|c} \hat{Q}_1 & 0 \\ \hline 0 & \hat{Q}_2 \end{array}\right] \left[\begin{array}{c} \tilde{u}_{1,k}^T Z^T + \Delta u_1 \\ \hline \tilde{U}_{2,k} + \Delta U_2 \\ \hline \tilde{v}_{1,k}^T + \Delta v_1 \\ \hline \tilde{V}_{2,k} + \Delta V_2 \end{array}\right] = \begin{bmatrix} \hat{u}_{1,k}^T \\ \hat{U}_{2,k} \\ \hat{v}_{1,k}^T \\ \hat{V}_{2,k} \end{bmatrix} = \hat{G}_k, \tag{7}$$

where for $m \doteq \max\{p-1, q-1\}$

$$\left\| \begin{bmatrix} \Delta u_1^T \\ \Delta U_2 \end{bmatrix} \right\|_F \leq 6m\epsilon \left\| \begin{bmatrix} \tilde{u}_{1,k}^T Z^T \\ \tilde{U}_{2,k} \end{bmatrix} \right\|_F \tag{8}$$

and

$$\left\| \begin{bmatrix} \Delta v_1^T \\ \Delta V_2 \end{bmatrix} \right\|_F \leq 6m\epsilon \left\| \begin{bmatrix} \tilde{v}_{1,k}^T Z^T \\ \tilde{V}_{2,k} \end{bmatrix} \right\|_F . \tag{9}$$

If we let

$$\Delta G_k = \begin{bmatrix} \Delta u_1 & \Delta U_2^T & \Delta v_1 & \Delta V_2^T \end{bmatrix}^T$$

then clearly

$$\|\Delta G_k\|_F \leq 6m\epsilon \|G_{k,Z}\|_F \leq 6m\epsilon \|G_k\|_F.$$

Also, if we let

$$\widehat{\Delta G_k} = \begin{bmatrix} \widehat{\Delta u} & \widehat{\Delta v} \end{bmatrix}^T$$

then the error bounds, (5) and (7), can be used to show that

$$\hat{G}_k^T \Sigma \hat{G}_k = (\tilde{G}_{k,Z} + \Delta G_k)^T \Sigma (\tilde{G}_{k,Z} + \Delta G_k)$$

and

$$(\tilde{G}_{k+1} + e_1 \widehat{\Delta u}^T)^T \Sigma (\tilde{G}_{k+1} + e_1 \widehat{\Delta u}^T) = (\hat{G}_k + e_{p+1} \widehat{\Delta v}^T)^T \Sigma (\hat{G}_k + e_{p+1} \widehat{\Delta v}^T),$$

where e_1 and e_{p+1} are standard basis vectors. This gives

$$\epsilon F_k = \tilde{G}_{k,Z}^T \Sigma \Delta G_k + \Delta G_k^T \Sigma \tilde{G}_{k,Z} - \begin{bmatrix} \tilde{u}_{1,k+1} & \hat{v}_{1,k} \end{bmatrix} \widehat{\Delta G_k} - \widehat{\Delta G_k}^T \begin{bmatrix} \tilde{u}_{1,k+1}^T \\ \hat{v}_{1,k}^T \end{bmatrix},$$

corresponding to a bound

$$\begin{aligned} \|\epsilon F_k\|_F &\leq 2\|\tilde{G}_{k,Z}\|_F \|\Delta G_k\|_F + 2(\|\hat{G}_k\|_F + \|\tilde{G}_{k+1}\|_F)\|\widehat{\Delta G_k}\|_F \\ &\leq 12m\epsilon \|\tilde{G}_{k,Z}\|_F^2 + 12.5\epsilon(\|\hat{G}_k\|_F + \|\tilde{G}_{k+1}\|_F)^2 \\ &\leq 12m\epsilon \|G_k\|_F^2 + 12.5\epsilon \left(\|G_k\|_F + \|G_{k+1}\|_F\right)^2 + O(\epsilon^2). \end{aligned}$$

Since Theorem 4.1 shows that

$$\|G_k\|_F^2, \|G_{k+1}\|_F^2 \leq 2\sqrt{k}\|A\|_F + \|G_1\|_F^2,$$

we get a bound on the local error of

$$\|\epsilon F_k\|_F \leq (50 + 12m)\epsilon(2\sqrt{k}\|A\|_F + \|G_1\|_F^2).$$

From (3), we see that

$$\|A - R^T R\|_F \leq (25 + 6m)(n-1)n\epsilon \left(2\sqrt{n}\|A\|_F + \|G_1\|_F^2\right). \tag{10}$$

5. An unstable implementation. In the last section the stability of the generalized Schur algorithm was proven for the case in which the hyperbolic transformations are applied in factored form. It has already been noted that the implementation is not unique. There are many transformations which introduce the needed zeros in the generators, and they can be applied in multiple ways: the obvious example is the application of the hyperbolic transformation in factored form, (4), versus the direct multiplication approach.

It turns out that the factored form of the hyperbolic transformation is crucial to the stability of the algorithm. To see this, take $\alpha = 4$ and the generators

$$G_0 = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} - \frac{1}{2} & \frac{1}{\sqrt{2}} - \frac{3}{2} & 1 \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} + \frac{1}{2} & \frac{1}{\sqrt{2}} + \frac{3}{2} \\ 0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 \\ 0 & 0 & 1-\eta & 1+2\sqrt{\eta} \end{bmatrix}. \tag{11}$$

For small η these generators correspond to an ill-conditioned A for which $\alpha = 4$. If the Schur algorithm with the hyperbolic transformations applied in factored form is used, then a small error $A - R^T R$ will be achieved, independent of the size of η. However, if the hyperbolic transformations are multiplied directly, then the backward error will depend to a significant extent on η.

To see why, note that this example gives, after two steps of the generalized Schur algorithm, the following generators:

$$G_Z = \begin{bmatrix} 1 & 1 \\ 0 & 2 \\ 1-\eta & 1+2\sqrt{\eta} \\ 0 & 1 \end{bmatrix}. \tag{12}$$

In fact, the example was obtained by performing a reversal of the algorithm on equation (12). The construction of the matrix in equation (11) was not really necessary since the essential mechanism of instability is represented in equation (12); this construction merely removes any objection as to whether the generators (12), corresponding to a displacement rank 2 matrix, could actually occur in the factorization of a displacement rank 4 matrix. If not, then one might argue that the instability shown when applying the Schur algorithm to equation (12) is not really relevant to the stability of the factorization of displacement rank 4 matrices. Since the matrix in equation (11) corresponds to four linearly independent generators which reduce to equation (11) in two steps of the Schur procedure, this objection can be dismissed.

Table 1 shows the dependence of the errors on η. It was necessary to go to very ill-conditioned matrices to demonstrate a significant loss of accuracy. The increase of the backward error by a factor 10^5 is more than seems reasonable in a backward stable algorithm, yet the increase in error is quite modest for an unstable algorithm in which the condition number is increased by a factor of 10^{10}. It turns out that this is primarily a result of the fact that instability is due to large local errors rather than unstable propagation of errors.

In contrast, over this very wide range of condition numbers, the error in the Cholesky factors computed by the stable form of the algorithm are all around 10^{-15}. This is consistent with the conclusion of the last section that the algorithm is stable.

It is interesting to note that the algorithm for $\alpha = 2$ is stable even when the hyperbolic transformations are applied directly. This is proven in [3]. The reason is

TABLE 1
Errors for different values of η.

η	$K(A)$	$\|A - CC^T\|$
10^{-3}	9.6×10^4	2.7×10^{-15}
10^{-8}	1×10^{10}	1.3×10^{-12}
10^{-13}	1×10^{15}	7×10^{-10}

that whenever a hyperbolic transformation of large norm which might magnify errors is performed, it can be proven that the norm of the generators drops drastically. Since the error bounds [4] for the unstabilized form of the hyperbolic transformation are proportional to the norm of the new generators multiplied by the norm of the transformation, the large norm of the hyperbolic transformation is canceled by a proportional decrease in the norm of the generators.

The reason for the decrease in norm of the generators is that whenever ρ is very close to negative one, the action of a hyperbolic transformation is close to just taking the difference between the two generators and scaling the result by $1/\sqrt{1-\rho^2}$. For the case $\alpha = 2$, whenever the leading nonzero elements are $O(\eta)$ apart, leading to a ρ very close to one, it can be proven that all the other components of the generators are within $O(\eta)$ of each other. As can be seen in the example here, this is clearly not the case for more than two generators: the leading elements from which the hyperbolic transformation is computed are within $O(\eta)$, but the other components of the two generators are only within $O(\sqrt{\eta})$. Their difference isn't small enough to completely cancel out the large norm of the hyperbolic transformation.

In Table 1, the backward error seems to display a proportionality to the square root of the condition number. This is a property which is suggested by Theorem 4.1 and the fact that the overall backward error is just a sum of local errors. These imply that neither the generators nor the effect of previous errors will be unduly magnified by the hyperbolic transformations. This makes it possible to concentrate our attention on just one stage of the algorithm. We do not expect the results in the overall factorization to be significantly worse than the worst possible loss of accuracy in a single stage.

Note that we have already proven that when applying a hyperbolic transformation the result is always a set of generators which satisfy the bound

$$\|G_k\|_F^2 \leq 2\sqrt{k-1}\|A\|_F + \|G_1\|_F^2$$

or, just looking at the two generators on which the hyperbolic transformation has acted,

$$\left\| \begin{bmatrix} u_{1,k}^T \\ v_{1,k}^T \end{bmatrix} \right\|_F^2 \leq 2\sqrt{k-1}\|A\|_F + \|G_1\|_F^2.$$

As a direct result of the analysis in [4], the local error introduced by the hyperbolic transformation which produces $u_{1,k}$ and $v_{1,k}$ will be proportional to the norm of the generators and the norm of the transformation, which can be bounded as follows:

$$\epsilon\|H_k\| \left\| \begin{bmatrix} u_{1,k}^T \\ v_{1,k}^T \end{bmatrix} \right\| \leq \frac{2\epsilon}{\sqrt{1-\rho^2}} \left\| \begin{bmatrix} u_{1,k}^T \\ v_{1,k}^T \end{bmatrix} \right\| \leq \frac{2\epsilon}{\sqrt{1-\rho^2}} (2\sqrt{k-1}\|A\|_F + \|G_1\|_F^2).$$

Although there does not appear to be a general theory relating the condition number of an arbitrary low α-rank matrix to the values ρ_k, it is proven in [5] that for a

positive-definite Toeplitz matrix

$$\prod_{k=1}^{n} \frac{1+\rho_k}{1-\rho_k^2} \leq \|T^{-1}\|_1 \leq \prod_{k=1}^{n-1} \frac{(1+|\rho_k|)^2}{1-\rho_k^2}.$$

If similar inequalities hold for matrices of displacement rank α, then this is enough to suggest that the errors will tend to be proportional to the square root of the condition number.

We can get a somewhat more conclusive result. If we assume that the factorization goes to completion, we can write of the leading elements of $\hat{u}_{11,k-1}$ and $\hat{v}_{11,k-1}$ from which ρ is computed,

$$\hat{v}_{11,k-1} = \hat{u}_{11,k-1}(1-\epsilon_1)$$

and $1 \geq \epsilon_1 \geq \epsilon$. We have assumed without loss of generality that $\hat{v}_{11,k-1} > 0$ and $\hat{u}_{11,k-1} > 0$. If this is not the case, then either generator can be multiplied by -1 without causing any problems.

This means that

$$\frac{1}{1-\rho^2} = \frac{\hat{u}_{11,k-1}^2}{\hat{u}_{11,k-1}^2 - \hat{v}_{11,k-1}^2} = \frac{1}{2\epsilon_1 - \epsilon_1^2} < \frac{1}{\epsilon_1} \leq \frac{1}{\epsilon}.$$

So the error incurred during the hyperbolic transformation will be at worst proportional to

$$2\sqrt{\epsilon}(2\sqrt{k-1}\|A\|_F + \|G_1\|_F^2).$$

This explains the results in Table 1: because of the stability of the error propagation in the algorithm, as well as the bound on the generators, the errors will be at worst a modest multiple of $\sqrt{\epsilon}$. Although the backward errors are dependent on the reflection coefficients, there is a limit to this dependence.

6. Summary. This paper has proven the stability of a method for Cholesky factorization of low α-rank matrices. The stability result is dependent on the particular implementation. Since the result depends on having a bound on the norm of the generators, it is important to choose transformations which permit such a bound. Working with transformations which can be factored as a product of a sequence of plane rotations and one hyperbolic transformation gives such a bound. In the case in which the algorithm is implemented with unfactored hyperbolic transformations, this bound on the generators is not sufficient for stability; however, bounds were given which support the assertion that the stability of the error propagation keeps the instability from being as bad as might otherwise be expected.

Although not discussed here, the approach in this paper can be used in a manner similar to that of [1] to provide a fast, weakly stable solution of Toeplitz least squares problems. The conclusions would be essentially the same as those in [1].

REFERENCES

[1] A. W. Bojanczyk, R. P. Brent, and F. R. De Hoog, *A weakly stable algorithm for general Toeplitz systems*, Numer. Algorithms, to appear.

[2] A. W. Bojanczyk, R. P. Brent, and F. R. De Hoog, *QR factorization of Toeplitz matrices*, Numer. Math., 49 (1986), pp. 81–94.

[3] A. W. Bojanczyk, R. P. Brent, F. R. De Hoog, and D. R. Sweet, *On the stability of the Bareiss and related Toeplitz factorization algorithms*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 40–57.

[4] A. W. Bojanczyk, R. P. Brent, P. Van Dooren, and F. R. De Hoog, *A note on downdating the Cholesky factorization*, SIAM J. Sci. Stat. Comput., 8 (1987), pp. 210–220.

[5] G. Cybenko, *The numerical stability of the Levinson-Durbin algorithm for Toeplitz systems of equations*, SIAM J. Sci. Stat. Comput., 1 (1980), pp. 303–319.

[6] K. A. Gallivan, S. Thirumalai, P. Van Dooren, and V. Vermaut, *High performance algorithms for Toeplitz and block Toeplitz matrices*, Linear Algebra Appl., 241–243 (1996), pp. 343–388.

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

[8] N. J. Higham, *Analysis of the Cholesky decomposition of a semi-definite matrix*, in Reliable Numerical Computation, M. G. Cox and S. J. Hammarling, eds., Oxford University Press, London, 1990, pp. 161–185.

[9] N. J. Higham, *Stability of the diagonal pivoting method with partial pivoting*, Numerical Analysis Report No. 265, Manchester Centre for Computational Mathematics, Manchester, England, July 1995.

[10] 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.

[11] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 119–123, January 1997

010

BOUNDS FOR THE DIFFERENCES OF MATRIX MEANS*

M. ALIĆ†, B. MOND‡, J. PEČARIĆ§, AND V. VOLENEC†

Abstract. The classical inequality between the weighted arithmetic and geometric means has recently been extended to means of positive definite matrices. Here we give bounds for the difference between such matrix means. Differences between other matrix means, such as the geometric and harmonic means, are also given. Finally, conditions for the reversal of the matrix inequalities obtained are also pointed out.

Key words. matrix means, arithmetic–geometric inequality

AMS subject classification. 15A45

PII. S0895479895290760

1. Introduction. The classical inequality between the weighted arithmetic and geometric means states that if $C_1, \dots, C_r$ and $w_1, \dots, w_r$ are positive real numbers, with $w_1 + \cdots w_r = 1$, then

$$G \equiv \prod_{i=1}^{r} C_i^{w_i} \le \sum_{i=1}^{r} w_i C_i \equiv A. \tag{1.1}$$

Equality holds in (1.1) if and only if $C_1 = \cdots = C_r$.

Alzer [1] (see also [2, p. 38]) proved the following lower bound for the difference between A and G:

$$0 \le Ae^{-G/A} - Ge^{-A/G} \le \frac{3}{e}(A - G). \tag{1.2}$$

The number $3/e$ cannot be replaced by a smaller number. Equality holds in (1.2) if and only if $C_1 = \cdots = C_r$.

In this note, we give a related result for the differences between arbitrary matrix means.

2. Preliminaries. Let $w_1, \dots, w_r$ be positive numbers such that $w_1 + \cdots + w_r = 1$ and let $C_1, \dots, C_r$ be $n \times n$ positive definite Hermitian matrices. Consider weighted power means of the matrices, defined by

$$M_s = (w_1 C_1^s + \cdots + w_r C_r^s)^{1/s}, \qquad s \ne 0, \tag{2.1}$$

$$\begin{aligned} M_0 = G = C_r^{1/2}(C_r^{-1/2} C_{r-1}^{1/2} \cdots (C_3^{-1/2} C_2^{1/2} (C_2^{-1/2} C_1 C_2^{-1/2})^{u_1} \\ \cdot C_2^{1/2} C_3^{-1/2})^{u_2} \cdots C_{r-1}^{1/2} C_r^{-1/2})^{u_{r-1}} C_r^{1/2}, \end{aligned}$$

where $u_i = 1 - w_{i+1} / \sum_{k=1}^{i+1} w_k$ for $i = 1, \dots, r-1$. We shall use the notation $A = M_1$ and $H = M_{-1}$ for arithmetic and harmonic means, respectively.

The following results were proved in [3].

* Received by the editors August 23, 1995; accepted for publication (in revised form) by T. Ando January 10, 1996.
http://www.siam.org/journals/simax/18-1/29076.html
† Department of Mathematics, University of Zagreb, Zagreb 10,000, Croatia.
‡ School of Mathematics, La Trobe University, Bundoora, Victoria 3083, Australia (b.mond@latrobe.edu.au).
§ Faculty of Textile Technology, University of Zagreb, Zagreb 10,000, Croatia.

THEOREM 2.1.

$$H \le G \le A. \tag{2.2}$$

In fact, (2.2) is the main result from [3] (see [4]). Various upper and lower bounds are also given. For example, the following inequalities hold [3].

THEOREM 2.2.

$$\begin{aligned} G\exp(G^{-1}A - I) &= \exp(AG^{-1} - I)G \\ &\ge (AG^{-1}A + G)/2 \\ &\ge A \ge G \ge H \\ &\ge 2(H^{-1}GH^{-1} + G^{-1})^{-1} \\ &\ge G\exp(I - H^{-1}G) = \exp(I - GH^{-1})G. \end{aligned} \tag{2.3}$$

Reverse inequalities for (2.2) are given in [4]. For such reverse results, it is necessary to change the requirement that all the weights w_i are positive.

THEOREM 2.3. *Let $w_i,\ i = 1, \dots, r$ be real numbers such that*

$$w_1 > 0,\ w_i < 0,\ i = 2, \dots, r; \quad w_1 + \cdots + w_r = 1. \tag{2.4}$$

Then

$$A \le G. \tag{2.5}$$

If we also have $w_1C_1^{-1} + \cdots + w_rC_r^{-1} > 0$, then

$$G \le H. \tag{2.6}$$

Equality holds in (2.5) *and* (2.6) *if and only if $C_1 = \cdots = C_r$.*

Finally we note the following result proved in [5].

THEOREM 2.4. *Let t, s be real numbers such that one of the following holds:*

(a) $t \ge s, \qquad t \notin (-1, 1), \qquad s \notin (-1, 1)$;

(b) $t \ge 1 \ge s \ge 1/2$; *or*

(c) $s \le -1 \le t \le -1/2$.

Then

$$M_t \ge M_s.$$

3. Main results. First we prove the following two lemmas.

LEMMA 3.1. *If $x \in (0, 1]$, then*

$$1 - x \le e^{-x} - xe^{-1/x} \le \frac{3}{e}(1 - x). \tag{3.1}$$

The values $3/e$ on the right and, correspondingly, 1 on the left cannot be replaced by smaller numbers. Equalities hold in (3.1) *if and only if $x = 1$.*

Proof. We shall modify the proof from [1]. For the proof of the second inequality, we shall, as in [1], consider the function

$$f(x) = \frac{3}{e}(1 - x) - e^{-x} + xe^{-1/x}.$$

Since $f''(x) = -e^{-x} + \frac{1}{x^3}e^{-1/x}$, we have (see [1]) that there exists a number $x_0 \in (0, 1)$ such that $f''(x) < 0$ for $x \in (0, x_0)$ and $f''(x) > 0$ for $x \in (x_0, 1)$. Namely, because

of $\lim_{x\to 0} f(x) = 3/e - 1 > 0$ and $f(1) = f'(1) = 0$, we get $f(x) \geq 0$ for all $x \in (0,1)$ with equality holding if and only if $x = 1$.

Here is a proof of the existence of such an x_0. Note that the function $h(x) = x - 1/x - 3\log x$ has the same sign as $f''(x)$. We have $x^3 h''(x) = 3x - 2$ which implies that h is strictly concave on $(0, 2/3)$ and strictly convex on $(2/3, 1]$. Since $h(1) = 0,\ h'(1) = -1$ and $\lim_{x\to 0} h(x) = -\infty$, there exists a number $x_0 \in (0, 2/3)$ such that $h(x) < 0$ for $x \in (0, x_0)$ and $h(x) > 0$ for $x \in (x_0, 1)$. Therefore, f is strictly concave on $(0, x_0)$ and strictly convex on $(x_0, 1)$, which completes the proof of the inequality $f(x) \geq 0$.

For the first part of the inequality, consider the function

$$k(x) = e^{-x} - xe^{-1/x} - 1 + x \qquad (0 \leq x \leq 1).$$

We have to prove that $k(x) \geq 0$ $(0 \leq x \leq 1)$. First, note that $k(1) = 0$ and $k(0) = 0$ while

$$k'(x) = -e^{-x} - e^{-1/x} - \frac{1}{x}e^{-1/x} + 1$$

so that $k'(1) = -3e^{-1} + 1 < 0$ and $k'(0) = 0$. It is sufficient to show that there is an $x_0,\ 0 < x_0 < 1$, such that

$$k'(x) > 0 (0 < x < x_0) \quad \text{and} \quad k'(x) < 0\ (x_0 < x < 1).$$

To this end, consider

$$k''(x) = e^{-x} - \frac{1}{x^3}e^{-1/x}.$$

The function $g(x) \equiv 3\log x - x + 1/x$ has the same sign as $k''(x)$. Since $-g(x) = h(x)$, the existence of x_0 satisfying $k'(x) > 0$ $(0 < x < x_0)$ and $k'(x) < 0$ $(x_0 < x < 1)$ is assured. This completes the proof.

That the constants $3/e$ on the right and 1 on the left cannot be replaced by smaller numbers can be seen from the fact that

$$\lim_{x\to 1} \frac{xe^{-1/x} - e^{-x}}{x - 1} = \frac{3}{e} \quad \text{and} \quad \lim_{x\to 0} \frac{xe^{-1/x} - e^{-x}}{x - 1} = 1.$$

LEMMA 3.2. *Let C and D be two $n \times n$ Hermitian matrices such that D is positive definite. Then*

$$D\exp(D^{-1}C) = D^{1/2}\exp(D^{-1/2}CD^{-1/2})D^{1/2} = \exp(CD^{-1})D. \tag{3.2}$$

Proof. We have

$$\begin{aligned} D\left(\sum_{i=0}^{\infty}(D^{-1}C)^i/i!\right) &= D^{1/2}\left(\sum_{i=0}^{\infty}(D^{1/2}(D^{-1}C)D^{-1/2})^i/i!\right)D^{1/2} \\ &= D^{1/2}\left(\sum_{i=0}^{\infty}(D^{-1/2}CD^{-1/2})^i/i!\right)D^{1/2} \\ &= D^{1/2}\left(\sum_{i=0}^{\infty}(D^{-1/2}(CD^{-1})D^{1/2})^i/i!\right)D^{1/2} \\ &= \left(\sum_{i=0}^{\infty}(CD^{-1})^i/i!\right)D. \end{aligned}$$

THEOREM 3.1. *Let C and D be two positive definite Hermitian matrices such that $C \geq D$. Then*

$$C - D \leq C\exp(-C^{-1}D) - D\exp(-D^{-1}C) \leq \frac{3}{e}(C - D). \tag{3.3}$$

The constants are the best possible and equality holds if and only if $C = D$.

Proof. Let M be a positive definite Hermitian matrix. If $M \leq I$, we have by Lemma 3.1,

$$I - M \leq \exp(-M) - M\exp(-M^{-1}) \leq \frac{3}{e}(I - M). \tag{3.4}$$

Since $C^{-1/2}DC^{-1/2} \leq I$, setting in (3.4), $M = C^{-1/2}DC^{-1/2}$, we have

$$\begin{aligned} I - C^{-1/2}DC^{-1/2} &\leq \exp(-C^{-1/2}DC^{-1/2}) - C^{-1/2}DC^{-1/2}\exp(-C^{1/2}D^{-1}C^{1/2}) \\ &\leq \frac{3}{e}(I - C^{-1/2}DC^{-1/2}). \end{aligned}$$

Premultiplication and postmultiplication by $C^{1/2}$ gives

$$\begin{aligned} C - D \leq C^{1/2}\exp(-C^{-1/2}DC^{-1/2})C^{1/2} \\ -DC^{-1/2}\exp(-C^{1/2}D^{-1}C^{1/2})C^{-1/2}C \leq \tfrac{3}{e}(C - D). \end{aligned}$$

Note that the first identity in (3.2), with C instead of D and $-D$ instead of C, is

$$C^{1/2}\exp(-C^{-1/2}DC^{-1/2})C^{1/2} = C\exp(-C^{-1}D),$$

while the second identity in (3.2), with C^{-1} instead of D and $-D^{-1}$ instead of C, is

$$C^{-1/2}\exp(-C^{1/2}D^{-1}C^{1/2})C^{-1/2} = \exp(-D^{-1}C)C^{-1}.$$

Thus, the last inequality gives

$$C - D \leq C\exp(-C^{-1}D) - D\exp(-D^{-1}C) \leq \frac{3}{e}(C - D),$$

which is (3.3).

Making use of the notation from section 2, we now give the following matrix version of (1.2) in Theorem 3.2.

THEOREM 3.2. *Let A and G be the arithmetic and geometric means of positive definite matrices $C_1, \ldots, C_r$. Then*

$$A - G \leq A\exp(-A^{-1}G) - G\exp(-G^{-1}A) \leq \frac{3}{e}(A - G). \tag{3.5}$$

The constants $3/e$ and 1 are best possible, and equality holds if and only if $C_1 = \cdots = C_r$.

Proof. This is a simple consequence of Theorems 3.1 and 2.1.

Another consequence of Theorems 3.1 and 2.1 is the following theorem.

THEOREM 3.3. *Let G and H be the geometric and harmonic means of r positive definite matrices C_i. Then*

$$G - H \leq G\exp(-G^{-1}H) - H\exp(-H^{-1}G) \leq \frac{3}{e}(G - H). \tag{3.6}$$

The constants $3/e$ *and* 1 *are the best possible, and equality holds if and only if* $C_1 = \cdots = C_r$.

A similar consequence of Theorems 3.1 and 2.4 is the following theorem.

THEOREM 3.4. *Let* $t,\ s$ *be real numbers such that one of conditions* (a), (b), *or* (c) *from Theorem* 2.4 *is satisfied. Then*

$$M_t - M_s \le M_t \exp(-M_t^{-1} M_s) - M_s \exp(-M_s^{-1} M_t) \le \frac{3}{e}(M_t - M_s). \tag{3.7}$$

The constants $3/e$ *and* 1 *are best possible, and equality holds if and only if* $C_1 = \cdots = C_r$.

For reversals of inequalities (3.5) and (3.6), we change the requirement that all the weights w_i must be positive. Then using Theorems 3.1 and 2.3 we obtain Theorem 3.5.

THEOREM 3.5. *Let* $w_i, i = 1, \ldots, r$ *be real numbers such that* $w_1 > 0,\ w_i < 0, i = 2, \ldots, r;\ w_1 + \cdots + w_r = 1$. *If*

$$w_1 C_1 + \cdots + w_r C_r > 0, \tag{3.8}$$

then

$$G - A \le G\exp(-G^{-1}A) - A\exp(-A^{-1}G) \le \frac{3}{e}(G - A). \tag{3.9}$$

If

$$w_1 C_1^{-1} + \cdots + w_r C_r^{-1} > 0, \tag{3.10}$$

then

$$H - G \le H\exp(-H^{-1}G) - G\exp(-G^{-1}H) \le \frac{3}{e}(H - G). \tag{3.11}$$

Moreover, if (3.8) *and* (3.10) *both hold, then*

$$H - A \le H\exp(-H^{-1}A) - A\exp(-A^{-1}H) \le \frac{3}{e}(H - A). \tag{3.12}$$

The constants $3/e$ *on the right and* 1 *on the left in* (3.9), (3.11), *and* (3.12) *are best possible, and equalities hold if and only if* $C_1 = \cdots = C_r$.

Acknowledgment. The authors wish to thank the referee for valuable suggestions that improved some of the inequalities in this paper.

REFERENCES

[1] H. ALZER, *A lower bound for the difference between the arithmetic and geometric means*, Nieuw Arch. Wisk., 8 (1990), pp. 195–197.

[2] D. S. MITRINOVIĆ, J. E. PEČARIĆ, AND A. M. FINK, *Classical and New Inequalities in Analysis*, Kluwer Academic Publishers, Norwell, MA, 1993.

[3] M. SAGAE AND K. TANABE, *Upper and lower bounds for the arithmetic-geometric-harmonic means of positive definite matrices*, Linear and Multilinear Algebra, 37 (1994), pp. 279–282.

[4] M. ALIĆ, B. MOND, J. E. PEČARIĆ, AND V. VOLENEC, *The arithmetic-geometric-harmonic means and related matrix inequalities*, Linear Algebra Appl., to appear.

[5] B. MOND AND J. E. PEČARIĆ, *A simple proof of generalized inequalities of Bhagwat and Subramanian*, Indian J. Math., 37 (1995).

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 124–139, January 1997

011

A CONSTRAINED PROCRUSTES PROBLEM*

LARS-ERIK ANDERSSON† AND TOMMY ELFVING†

Abstract. The following constrained matrix problem is studied. Find the matrix X that minimizes the Frobenius norm of $AX - B$, with A and B as given matrices and where X belongs to a closed convex cone. In particular we consider the cone of symmetric positive semidefinite (SPSD) matrices and the cone of (symmetric) elementwise nonnegative matrices. The optimal matrix is characterized, and the results are specialized to the two cases above. Further, we report from a numerical study of some projection-type algorithms.

Key words. constrained matrices, convex cones, positive semidefinite, Procrustes

AMS subject classifications. 65F20, 65F30, 65K10

PII. S0895479894277545

1. Introduction.

Let H be the linear space of real matrices with a fixed dimension. On H we introduce the inner product $\langle \cdot, \cdot \rangle$,

$$\langle X, Y \rangle = \text{trace}\,(XY^T), \qquad X, Y \in H. \tag{1.1}$$

Let $f(X)$ be a convex function on H and let C be a closed convex and nonempty cone in H. The function f is called coercive on C if

$$\lim_{||X|| \to \infty, X \in C} f(X) = \infty.$$

We shall study the problem

$$\min_{X \in C} f(X). \tag{1.2}$$

We first have the well-known result.

LEMMA 1.1. *Assume that f is convex and coercive. Then problem* (1.2) *has a solution. If in addition f is strictly convex, the solution is unique.*

In the next section we will give a short derivation of the optimality conditions for problem (1.2). In section 3 we consider the quadratic functional

$$f(X) = \|AX - B\|_F^2, \tag{1.3}$$

which yields the *constrained Procrustes problem* (using the notation by Higham [18])

$$\text{CPP:} \quad \min_{X \in C} \|AX - B\|_F^2. \tag{1.4}$$

Here A is an $m \times n$ matrix, with $m \geq n$, X an $n \times q$ matrix, and B an $m \times q$ matrix. The index F denotes the Frobenius norm of a matrix, i.e.,

$$\|Z\|_F^2 = \sum_{i,j} z_{ij}^2 = \text{trace}\,(ZZ^T) = \text{trace}\,(Z^T Z). \tag{1.5}$$

* Received by the editors November 23, 1994; accepted for publication (in revised form) by N. J. Higham January 29, 1996. The work of the second author was supported by the Swedish Research Council for Engineering Sciences (TFR).

http://www.siam.org/journals/simax/18-1/27754.html

† Department of Mathematics, Linköping University, S-581 83 Linköping, Sweden (leand@math.liu.se, toelf@math.liu.se).

If $\text{rank}(A) = n$ then, by Lemma 1.1, CPP has a unique solution.

A special case is the *matrix nearness problem*

$$\min_{X \in C} \|X - B\|_F^2. \tag{1.6}$$

We denote the solution of the matrix nearness problem (1.6) as

$$P_C(B) = \arg\min_{X \in C} \|X - B\|_F^2. \tag{1.7}$$

Here we study two types of constraints: elementwise nonnegativity and definiteness. First let

$$C = C_1 = \{X \in H : X \geq 0\}, \tag{1.8}$$

where $X \geq 0$ means $x_{ij} \geq 0 \ \forall \ i, j$. Of course one may also consider nonnegativity for a subset of the elements of X. We also study nonnegativity in the symmetric case $(q = n)$,

$$C = C_2 = \{X \in R^{n \times n} : X = X^T, X \geq 0\}. \tag{1.9}$$

Next we consider the symmetric positive semidefinite (SPSD) case

$$C = C_3 = \{X \in R^{n \times n} : X = X^T, z^T X z \geq 0 \ \forall \ z \in R^n\}. \tag{1.10}$$

As the final example consider

$$C = C_4 = \{X \in R^{m \times n} : \text{sparsity}(X) = \text{sparsity}(Y)\}, \tag{1.11}$$

where Y is a given matrix and $\text{sparsity}(Y)$ denotes its sparsity pattern; i.e., $\text{sparsity}(Y)$ is an $m \times n$ matrix with elements equal to zero if the corresponding elements in Y are zero and otherwise equal to one.

Obviously $P_{C_4}(B) = B \odot \text{sparsity}(Y)$, where $\odot$ denotes elementwise multiplication. For the matrix nearness problem (1.7) it is an easy exercise to verify that

$$P_{C_1}(B) = \max(0, B) \tag{1.12}$$

and

$$P_{C_2}(B) = \max(0, (B + B^T)/2). \tag{1.13}$$

Here $\max(0, B)$ means elementwise maximization.

In [19] Higham studied the matrix nearness problem for the SPSD case. In order to state one of his results we will introduce the spectral decomposition of a symmetric matrix Z,

$$Z = U^T D U, \tag{1.14}$$

with

$$D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n), \qquad UU^T = U^T U = I. \tag{1.15}$$

Define

$$D^+ = \text{diag}(\lambda_1^+, \lambda_2^+, \ldots, \lambda_n^+), \tag{1.16}$$

where

$$\lambda_i^+ = \max(0, \lambda_i) \tag{1.17}$$

and

$$Z^+ = U^T D^+ U. \tag{1.18}$$

A special case of Theorem 2.4 in [19] shows that

$$P_{C_3}(B) = (B + B^T)^+/2. \tag{1.19}$$

For the case $A \neq I$ there does not exist any explicit expression for the solution (for $C = C_1$, C_2, and C_3). However, for the SPSD case, one can always reduce the general case to the case when A is a diagonal matrix—Theorem 3.3. Also for the case when $A^T A$ commutes with the matrix $G = A^T B + B^T A$, an expression analogous to (1.19) can be found in Theorem 3.4.

We also provide a simple example showing that, for the SPSD case, when $\operatorname{rank}(A) < n$, CPP may have no solution.

In section 4 we present some numerical results. Here we test three different projection-type methods: the gradient projection algorithm and a method due to Han and Lou [17]. We also propose a variant of the method of parallel tangents (PARTAN) [22, Chapter 8.7] and consider its convergence properties. We observe that the methods are sensitive to the conditioning of the matrix A. Also in some cases it is hard to find a good value of the steplength parameter.

We end this section by giving a few more references. In [9], [15] some early investigations are made. Halmos [15] reformulates the matrix nearness problem (1.6) (using the Euclidean norm) with $C = C_3$ as a minimization problem in only one (nonnegative) variable. Brock [7] considers CPP with $C = C_3$. However, as pointed out in [18], his method of solution does not take definiteness into account. The symmetric Procrustes problem arises, for example, in the investigation of elastic structures; see [27, Chapter 3.5]. In a recent contribution Suffridge and Hayden [28] solved the matrix nearness problem for a Hermitian positive definite Toeplitz matrix.

Allwright in [1] considers problem (1.4) for $C = C_3$. He also provides an algorithm based on a special characterization of C_3. In [29] Woodgate treats the same problem. He also provides an algorithmic scheme but provides no numerical results. We also mention [8] where among other things the problem of finding a least squares approximation, in the Frobenius norm, to a given symmetric matrix with a prescribed set of eigenvalues is treated. In [2] the same problem is considered but with the additional constraint that the sparsity pattern of the solution is fixed. We note that the two last examples do not give rise to convex constraints. (The set of eigenvalues is not preserved under addition!)

Ruhe [25] considers the matrix nearness problem when X is normal, i.e., $XX^T = X^T X$. In Golub and Van Loan [13] the orthogonal Procrustes problem $q = n$, $X^T X = I$ is described. Again the last two problems do not fit into the framework considered here since the set of normal matrices and the set of orthogonal matrices are not convex. As a general reference to the area of nonnegative matrices we refer to the book by Berman and Plemmons [3] and for least squares problems to the monograph by Björck [6] and the book by Golub and Van Loan [13].

2. Characterization of an optimal matrix. Here and in what follows (unless otherwise stated) we will assume that f is convex and that $f \in C^1$. Also we assume that problem (1.2) has a solution. This is fulfilled, for example, if f is coercive. Let ∇f be the gradient of f and ϵ a real number. Then

$$f(X + \epsilon E) = f(X) + \epsilon\langle \nabla f(X), E\rangle + o(\epsilon), \tag{2.1}$$

where, from convexity, $o(\epsilon) \geq 0$.

We first give the well-known characterization of the optimal solution of problem (1.2); see, e.g., [4, Proposition 3.1, p. 201] or [24, Theorem 27.4].

LEMMA 2.1. *Let C be a closed convex set in H. A necessary and sufficient condition for a matrix $X \in C$ to be a solution of* (1.2) *is that*

$$X + E \in C \Longrightarrow \operatorname{trace}(\nabla f(X)E^T) \geq 0. \tag{2.2}$$

Proof. First suppose that X solves (1.2) and that $X + E \in C$. Then $\lambda X + (1 - \lambda)(X + E) = X + (1 - \lambda)E \in C$ for all $\lambda \in [0, 1]$. By (2.1), with $\epsilon = 1 - \lambda$, it follows that $\epsilon\langle \nabla f(X), E\rangle \geq 0$ for ϵ small enough. Consequently $\operatorname{trace}(\nabla f(X)E^T) \geq 0$.

Next, assume that the implication (2.2) is valid. Then it follows from (2.1) with $\epsilon = 1$ that

$$X + E \in C \Longrightarrow f(X + E) \geq f(X),$$

i.e., that X minimizes f. ∎

Let C^* be the dual cone of C,

$$C^* = \{Y \in H : X \in C \Rightarrow \langle X, Y\rangle \geq 0\}.$$

Note that this concept is well defined for C being any set in H (cf. [26, section 2.7]), and in any case C^* is a convex cone. However, to derive the results below we must specialize to the case when C is a closed convex cone (as in problem (1.2)).

LEMMA 2.2. *Let C be a closed convex cone in H. If X solves* (1.2) *then $\nabla f(X) \in C^*$.*

Proof. Pick any $E \in C$. Then $(X + E)/2 \in C$, i.e., $X + E \in C$. By Lemma 2.1,

$$\operatorname{trace}(\nabla f(X)E^T) = \langle \nabla f(X), E\rangle \geq 0;$$

i.e., $\nabla f(X) \in C^*$. ∎

LEMMA 2.3. *Let C be a closed convex cone in H. If X is a solution of* (1.2) *then* $\operatorname{trace}(\nabla f(X)X^T) = 0$.

Proof. Pick $E = \epsilon X$, $\epsilon \in \mathbf{R}$. Then $X + E = (1 + \epsilon)X \in C$ for $\epsilon > -1$. By Lemma 2.1 it follows that

$$\operatorname{trace}(\nabla f(X)E^T) = \epsilon \operatorname{trace}(\nabla f(X)X^T) \geq 0.$$

Hence $\operatorname{trace}(\nabla f(X)X^T) = 0$. ∎

THEOREM 2.4. *Let C be a closed convex nonempty cone in H. Then X is a solution of* (1.2) *if and only if*

$$X \in C, \quad \nabla f(X) \in C^*, \quad \text{and} \quad \operatorname{trace}(\nabla f(X)X^T) = 0.$$

Proof. First assume that X solves (1.2). Then $\nabla f(X) \in C^*$ and $\operatorname{trace}(\nabla f(X)X^T) = 0$ by Lemmas 2.2 and 2.3, respectively. Conversely, assume that $X \in C$, $\nabla f(X) \in C^*$, and $\operatorname{trace}(\nabla f(X)X^T) = 0$. Let E be such that $X + E \in C$. Then

$$\begin{aligned}\operatorname{trace}(\nabla f(X)E^T) &= \operatorname{trace}(\nabla f(X)(X+E)^T) - \operatorname{trace}(\nabla f(X)X^T)\\ &= \operatorname{trace}(\nabla f(X)(X+E)^T).\end{aligned}$$

But $\nabla f(X) \in C^*$ and $X + E \in C$ implies that $\langle \nabla f(X), X + E\rangle \geq 0$. Hence the result follows from Lemma 2.1. □

REMARK 2.1. *This characterization may also be derived from the Fenchel duality theory*; *see section* 31 *in Rockafellar* [24].

REMARK 2.2. *Let X_0 be a fixed matrix. The slightly generalized problem*

$$\min_{X - X_0 \in C} f(X)$$

can be handled by simply putting

$$Y = X - X_0 \quad \textit{and} \quad g(Y) = f(Y + X_0).$$

The optimality conditions then become

$$X - X_0 \in C, \quad \nabla f(X) \in C^*, \quad \operatorname{trace}(\nabla f(X)(X - X_0)^T) = 0.$$

REMARK 2.3. *Assume that $C = \cap_1^p B_i$, with B_i closed and convex cones, such that $\cap_1^p ri(B_i) \neq \emptyset$, where $ri(B_i)$ denotes the relative interior of B_i; see* [24, p. 44]. *Then see, e.g.,* [24, *Corollary* 16.4.2]:

$$C^* = \sum_1^p B_i^*.$$

3. The Procrustes problem. One may verify that

$$\begin{aligned}&f(X) = \|AX - B\|_F^2\\ &= \operatorname{trace}(X^TA^TAX) - 2\operatorname{trace}(B^TAX) + \operatorname{trace}(B^TB),\end{aligned} \tag{3.1}$$

and hence

$$\begin{aligned}&f(X + \epsilon E)\\ &= f(X) + 2\epsilon \operatorname{trace}\{E^T(A^TAX - A^TB)\} + \epsilon^2 \operatorname{trace}(E^TA^TAE).\end{aligned} \tag{3.2}$$

It follows that

$$\frac{1}{2}\nabla f(X) = A^T(AX - B) \tag{3.3}$$

and that the quadratic form $\frac{1}{2}\nabla^2 f(X)(E, E)$ is represented by

$$\frac{1}{2}\nabla^2 f(X)(E,E) = \operatorname{trace}(E^TA^TAE) = \langle E, A^TAE\rangle. \tag{3.4}$$

Here $\nabla^2 f$ is the Hessian of f.

In Hall [14, pp. 353–354] it is shown that $C_1^* \subset C_1$ and it is an easy exercise to verify that also $C_1 \subset C_1^*$. Hence for the case of elementwise nonnegativity (without symmetry), we get by Theorem 2.4 the following theorem.

THEOREM 3.1. *Let $C = C_1$. Then X is a solution of CPP if and only if*

$$X \geq 0, \quad A^T(AX - B) \geq 0, \quad \textit{and} \tag{3.5}$$
$$\operatorname{trace}(A^T(AX - B)X^T) = 0.$$

These formulas correspond to the Karush–Kuhn–Tucker conditions, with the Lagrange parameter being eliminated; see, e.g., [10, p. 200]. Note also that for $q = 1$, i.e., for the case when X and B are columns, we recover the nonnegatively constrained least squares problem; see, e.g., [6]. Then the preceding conditions become

$$x \geq 0, \quad A^T(Ax - b) \geq 0, \quad x_i(A^T(Ax - b))_i = 0, \quad i = 1, 2, \ldots, n.$$

We now consider the symmetric case. Let $E = E^T$ be an $n \times n$ matrix. Then

$$\operatorname{trace}(E^T A^T B) = \operatorname{trace}(B^T AE) = \operatorname{trace}(E^T B^T A).$$

Similarly

$$\operatorname{trace}(E^T A^T AX) = \operatorname{trace}(X^T A^T AE) = \operatorname{trace}(E^T XA^T A)$$

and it follows, using (3.2) and (3.3), that

$$\langle \nabla f(X), E \rangle = \operatorname{trace}(E^T(A^T AX + XA^T A - A^T B - B^T A));$$

i.e.,

$$\langle \nabla f(X), E \rangle = \langle A^T AX + XA^T A - A^T B - B^T A, E \rangle.$$

We conclude that

$$S(X) := \nabla f(X) = XA^T A + A^T AX - G, \tag{3.6}$$

$$G = A^T B + B^T A. \tag{3.7}$$

We sometimes write S instead of $S(X)$. For the case $C = C_2$ the conditions corresponding to (3.5) become (using that S given by (3.6) is symmetric)

$$X = X^T \geq 0, \quad S \geq 0, \quad \text{and} \quad \operatorname{trace}(SX) = 0.$$

We next look at the definite case and first remark that any solution of the normal equations (a special case of the Sylvester equations)

$$S(X) = 0 \tag{3.8}$$

yields a solution of the symmetric Procrustes problem (i.e., only requiring symmetry of X); see Higham [18]. Higham also notes that for G positive semidefinite the solution of (3.8) is also definite. This observation is generalized by Woodgate for $A \neq I$; see Theorem 2.4 in [29]. In the general case we have the following characterization.

THEOREM 3.2. *Let $C = C_3$. X is then a solution of CPP if and only if*

$$X \textit{ is SPSD}, \quad S(X) \textit{ is SPSD}, \quad \textit{and} \quad S(X)X = 0. \tag{3.9}$$

Proof. We will derive the result from Theorem 2.4 and first verify that

$$\operatorname{trace}(SX) = 0 \Longleftrightarrow SX = 0.$$

Let $S = L^T L$, $X = R^T R$. Then

$$\operatorname{trace}(SX) = \operatorname{trace}(L^T L R^T R) = \operatorname{trace}(L R^T R L^T) = \|L R^T\|_F^2 = 0$$

implies that

$$L R^T = 0 \quad \text{and} \quad SX = L^T L R^T R = 0.$$

We next show, under the assumption that H is the space of symmetric matrices, that $C_3 = C_3^*$. (In [14, pp. 353–354] it is shown that $C_3^* \subset C_3$.) Note first by (3.6) that S is symmetric. Let $Y \in C_3^*$ be symmetric. Pick $X = xx^T \in C_3$. Then

$$\langle X, Y \rangle = \operatorname{trace}(XY) = \operatorname{trace}(xx^T Y) = \operatorname{trace}(x^T Y x) \geq 0,$$

which shows that $Y \in C_3$ (here we use that $S = S^T$).

On the other hand let $Y \in C_3$. Pick $X \in C_3$. We may write $Y = L^T L$, $X = R^T R$. It follows, as above, that

$$\langle X, Y \rangle = \operatorname{trace}(XY) = \|L R^T\|_F^2 \geq 0;$$

i.e., $Y \in C_3^*$. ∎

REMARK 3.1. *Woodgate, Theorem* 2.3 *in* [29]*, has obtained essentially the same result. One difference is that he gives the condition* $\operatorname{trace}(SX) = 0$ *rather than* $SX = 0$. *Woodgate does not consider general cone constraints but only* $C = C_3$.

REMARK 3.2. *It is an easy exercise to verify that conditions* (3.9) *hold for the solution* (1.19) *of the matrix nearness problem associated with* $C = C_3$.

REMARK 3.3. *If* X *solves CPP, with* $C = C_3$, *then in some suitable orthonormal basis,* X *and* S *have the following block structure:*

$$X = \begin{pmatrix} X_{11} & 0 \\ 0 & 0 \end{pmatrix} \quad \textit{and} \quad S = \begin{pmatrix} 0 & 0 \\ 0 & S_{22} \end{pmatrix},$$

where X_{11} *and* S_{22} *are square matrices of dimension* $n - k$ *and* k, $0 \leq k \leq n$, *respectively.*

Let

$$A = P \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} Q^T \tag{3.10}$$

with P, $m \times m$, Σ, $n \times n$ diagonal, and Q, $n \times n$ be the singular value decomposition (SVD) of A (i.e., P, Q are orthogonal matrices). The Frobenius norm is invariant under orthogonal transformations. Hence for the case $q = n$, as in [18],

$$\|AX - B\|_F^2 = \|\Sigma Y - B_1\|_F^2 + \|B_2\|_F^2. \tag{3.11}$$

Here

$$Y = Q^T X Q, \qquad P^T B Q = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}. \tag{3.12}$$

Since the cone of SPSD matrices is invariant under a congruence transformation, the following result is valid.

THEOREM 3.3. *Let $C = C_3$. Then X is a solution of CPP if and only if $X = QYQ^T$ and Y is the solution of*

$$\min_{Y \in C_3} \|\Sigma Y - B_1\|_F^2.$$

As stated previously, we always assume that the CPP has a solution. This is fulfilled, e.g., when f is coercive. Now coercivity is equivalent, as is easily seen, with the condition rank$A = n$. This condition is too restrictive, however, and in fact Woodgate gives necessary and sufficient conditions for the existence of a solution ($C = C_3$) when rank$(A) < n$; see Theorem 2.2 in [29]. We also remark that Gowda [12] gives a sufficient condition (weaker than coercivity) to ensure the existence of a solution to CPP (i.e., not only for $C = C_3$).

We provide the following simple example showing that when rank $(A) < n$, $C = C_3$ there may be no solution. Note that the normal equations (3.8) have a solution also when rank $(A) < n$; see, e.g., [18].

EXAMPLE 3.1. *Let $m = n = 2$ and*

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \text{and} \quad X = \begin{bmatrix} x_{11} & x_{12} \\ x_{12} & x_{22} \end{bmatrix}.$$

Then

$$f(X) = \|AX - B\|_F^2 = x_{11}^2 + (x_{12} - 1)^2$$

is to be minimized under the condition that X is positive semidefinite, i.e., that

$$x_{11} \geq 0 \quad \text{and} \quad x_{11}x_{22} - x_{12}^2 \geq 0.$$

Note that for the unconstrained problem, $x_{11} = 0$, $x_{12} = 1$ yields the symmetric (but indefinite) solution. By selecting a sequence

$$X_k = \begin{bmatrix} 1/k & 1 \\ 1 & k \end{bmatrix},$$

it is seen that $X_k \in C_3$ and that

$$f(X_k) = 1/k^2 \to 0 \quad \text{as} \quad k \to \infty.$$

The value 0, however, is not taken for any matrix $X \in C_3$. Hence CPP has no solution in this case.

We finally comment on a special case of the symmetric positive definite case.

THEOREM 3.4. *Let $C = C_3$. Suppose that $A^TAG = GA^TA$ and that A has full rank. Then for every solution X_U of the unconstrained problem $S(X) = 0$, the matrix $X = X_U^+$ solves the constrained problem.*

Proof. By a well-known result A^TA and G have a common set of eigenvectors, i.e. (using the notation in (3.10)),

$$A^TA = Q\Sigma^2Q^T, \quad G = Q\Sigma_GQ^T, \quad \Sigma_G = \operatorname{diag}(\bar{\sigma}_i).$$

Now, again using the SVD of A,

$$G = A^TB + B^TA = Q \begin{bmatrix} \Sigma & 0 \end{bmatrix} P^TB + B^TP \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} Q^T.$$

It follows, using (3.12),

$$\Sigma_G = Q^T G Q = \begin{bmatrix} \Sigma & 0 \end{bmatrix} P^T B Q + Q^T B^T P \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} = \Sigma B_1 + B_1^T \Sigma.$$

By the argument preceding Theorem 3.3, $X_U = QY^TQ^T$ solves the unconstrained problem if and only if

$$S(Y) = Y\Sigma^2 + \Sigma^2 Y - \Sigma B_1 - B_1^T \Sigma = Y\Sigma^2 + \Sigma^2 Y - \Sigma_G = 0,$$

i.e., if and only if

$$y_{ij}(\sigma_i^2 + \sigma_j^2) = 0, \qquad i \neq j.$$

Here y_{ij} denote the elements of the matrix Y and $\Sigma = \mathrm{diag}(\sigma_i)$. Since A has full rank we have $\sigma_i > 0$ for all i and we conclude that Y is diagonal and moreover that $Y = \Sigma^{-2}\Sigma_G/2$. Now take $X = Q^T Y^+ Q$, with $Y^+ = \Sigma^{-2}(\Sigma_G)^+/2$. It remains to verify that

(1) $S(Y^+)$ is positive semidefinite.
(2) $S(Y^+)Y^+ = 0$.

We show (2) and leave (1) as an easy exercise:

$$\begin{aligned} S(Y^+)Y^+ &= Y^+\Sigma^2 Y^+ + \Sigma^2 {Y^+}^2 - \Sigma_G Y^+ \\ &= \frac{1}{4}\Sigma_G^{+2}\Sigma^{-2} + \frac{1}{4}\Sigma_G^{+2}\Sigma^{-2} - \frac{1}{2}\Sigma_G\Sigma_G^+\Sigma^{-2} = 0. \end{aligned}$$ □

REMARK 3.4. *The condition $A^TAG = GA^TA$ is satisfied if $A = I$. Then $X_U = G/2 = (B + B^T)/2$ and Higham's solution $X = (B + B^T)^+/2$ is recovered.*

4. Numerical results. We may recognize at least three different ways to construct a numerical algorithm for CPP. One approach is to transform the problem. For example, for the case $C = C_3$ we may substitute $X = LL^T$. A drawback with this approach is that the transformed problem then no longer is quadratic. Another way is to try to satisfy the optimality conditions in Theorem 2.4 by applying some type of projection procedure. The third way is to work directly with CPP formulation. Here we will only consider the last approach and further restrict our numerical study to certain projection-type algorithms. Of course other classes of algorithms may well be worth pursuing, e.g., active set methods; see, e.g., [10, Chapter 10]. We mention in particular two methods which both seem well suited for solving CPP for $C = C_1$. The algorithm by Björck [5] is of active set type and is specially designed for least squares with bound constraints. The recent algorithm by Friedlander and Martinez [11] is for bound constraint semidefinite quadratic problems. This method is a combination of the gradient projection and the conjugate gradient method.

The first method we consider here is the (by now classical) gradient projection algorithm, originally devised by Goldstein, Levitin, and Polyak. The method is

$$X^{k+1} = P_C(X^k - \gamma \nabla f(X^k)). \tag{4.1}$$

We refer to the book by Bertsekas and Tsitsiklis [4] for a nice presentation of the method. In particular Proposition 3.4, p. 214 in [4] shows that (4.1) is convergent for $0 < \gamma < 1/\lambda_{max}(A^TA)$ when applied to CPP. Formulas (1.12), (1.13), and (1.19),

respectively, are used to compute the projection operator in (4.1). The steplength γ was taken as

$$\gamma = 1/(\lambda_{max}(A^TA) + \lambda_{min}(A^TA)) \tag{4.2}$$

following the theory for unconstrained optimization; see, e.g., [20, Theorem 1, p. 84]. Here $\lambda_{max}(S)$ and $\lambda_{min}(S)$ denote the largest and smallest eigenvalues of a symmetric matrix S, respectively.

The second method we tested was proposed by Han and Lou [17] for minimizing a convex function over the intersection of convex sets (here there is only one convex set). For simplicity we now only formulate their method for the symmetric case ($C = C_2$, C_3) and also assuming that $\text{rank}(A) = n$. Then it is easy to verify that S^{-1} exists. Let $S^k = \nabla f(X^k)$. It can be shown that the method becomes

$$X^{k+1} = S^{-1}\{(1/\alpha)P_C(X^k - \alpha S^k) - (1/\alpha)(X^k - \alpha S^k)\}. \tag{4.3}$$

Convergence is assured for $\alpha > 1/\lambda_{min}(A^TA)$; see [17].

There are also two related schemes, one by Han [16] and the other by Iusem and De Pierro [21]. However these two schemes utilize weighted projections for which there are no explicit solutions of the projection problem (1.6). The same remark also holds for the scaled gradient projection method [4].

The third and last method we tested is based on the so-called method of parallel tangents; see [22, Chapter 8.7]. This method was devised for unconstrained optimization. Given X^0, X^1, a typical step consists of

$$\begin{aligned} Z^{k+1} &= X^k - \gamma\nabla f(X^k), \\ X^{k+1} &= X^{k-1} + \alpha_k(Z^{k+1} - X^{k-1}). \end{aligned} \tag{4.4}$$

For our problem we propose the following modification (with given $X^0, X^1 \in C$, and $f(X^1) < f(X^0)$):

$$Z^{k+1} = P_C(X^k - \gamma\nabla f(X^k)), \tag{4.5}$$

$$X^{k+1} = X^{k-1} + \alpha_k(Z^{k+1} - X^{k-1}). \tag{4.6}$$

Here

$$0 < \gamma < 2/L, \tag{4.7}$$

with L the Lipschitz constant of f. For f given by (1.3), $L = 2\lambda_{max}(A^TA)$. The steplength α_k should be taken such that

$$f(X^{k+1}) \le f(Z^{k+1}), \qquad X^{k+1} \in C. \tag{4.8}$$

Note that with $\alpha_k = 1$ we retrieve the gradient projection method. Note also, assuming $X^{k-1} \in C$, that $X^{k+1} \in C$ for $\alpha_k \in [0, 1]$.

In the practical application of the method the following procedure for picking α_k was used. First,

$$\alpha_k := 1 \text{ for } k = j(n+1), \qquad j = 1, 2, \ldots; \tag{4.9}$$

i.e., the method was restarted every $(n+1)$th step. For other values of k we first compute the number $\bar{\alpha}_k$ by an unconstrained linesearch in f and get the formula

$$\bar{\alpha}_k = -\frac{\langle AX^{k-1} - B, A(Z^{k+1} - X^{k-1})\rangle}{||A(Z^{k+1} - X^{k-1})||_F^2}.$$

Then the following rule for picking α_k was used:

$$\alpha_k = \begin{cases} \bar{\alpha}_k & \text{if } 0 \le \bar{\alpha}_k \le \alpha_{max}, \\ 1 & \text{else.} \end{cases}$$

Here α_{max} is estimated heuristically such that conditions (4.8) are fulfilled. Other possibilities for picking α_k, which we have not investigated here, include exact or approximate constrained linesearch, possibly using merit functions other than f. For a discussion of barrier and penalty methods, see, e.g., Chapter 12 in [10].

We realize that method (4.5)–(4.6) can be seen as a gradient projection method extended with a so-called spacer step; see, e.g., Luenberger [22, Chapter 7.10]. In the spacer step theorem, given in Luenberger [22], a point-to-set-type of algorithm is considered. Below we give another proof but only for an algorithm of point-to-point-type. On the other hand we explicitly take into account constraints. Conditions are also given such that the whole sequence, rather than a subsequence, converges.

LEMMA 4.1. *Let* $\Gamma = \{x : f(x) = \min_{y \in C} f(y)\} \neq \emptyset$, *with* f, *be lower semicontinuous and let* B *be a continuous function* $B : C \to C$. *Further let* K *be an infinite index set and* N *the set of natural numbers. Assume the following:*

(i) $y = B(x),\ x \notin \Gamma \Rightarrow f(y) < f(x)$,
(ii) $x^{k+1} = B(x^k)\ \forall k \in K$,
(iii) $f(x^{k+1}) \le f(x^k)\ \forall k \in N$,
(iv) $\{x : f(x) \le f(x^0)\}$ *is compact.*

Then $\{x^k\}_{k \in K}$ *contains at least one convergent subsequence and for any such convergent subsequence* $\{x^{k_\nu}\}_{\nu=1}^{\infty}$ *it holds* $x^{k_\nu} \to x^* \in \Gamma$. *If in addition* $f \in C^1$ *is strictly convex and* $x^k \in C\ \forall k \in N$, *then the whole sequence converges toward the unique* $x^* \in \Gamma$.

Proof. By (iii) and (iv) there is a subsequence $\{x^{k_\nu}\},\ k_\nu \in K$ such that $x^{k_\nu} \to x^*$ and $f(x^{k_\nu}) \to f(x^*)$. Now $k_{\nu+1} \ge k_\nu + 1$ implies by (iii) that $f(x^{k_{\nu+1}}) \le f(x^{k_\nu+1}) \le f(x^{k_\nu})$; hence $f(x^{k_\nu+1}) \to f(x^*)$. But $x^{k_\nu+1} = B(x^{k_\nu}) \to B(x^*)$, so $f(x^{k_\nu+1}) \to f(B(x^*))$. Thus $f(x^*) = f(B(x^*))$ and it follows using (i) that $x^* \in \Gamma$.

By (iii) and (iv), $f(x^k) \to f(x^*)$. Now if $x^k \in C$ then $f(x^* + \epsilon(x^k - x^*)) \ge f(x^*),\ \ 0 < \epsilon < 1$. Hence $\frac{d}{d\epsilon} f(\cdot)|_{\epsilon=0} = \langle \nabla f(x^*), x^k - x^* \rangle \ge 0$. By Taylor's formula we get

$$f(x^k) - f(x^*) = \langle \nabla f(x^*), x^k - x^* \rangle + \varphi(x^k - x^*),$$

where by assumption the function φ is strictly convex and nonnegative. Letting $k \to \infty$ we have $\varphi(x^k - x^*) \to 0$, and from the assumed strict convexity of f it follows that $x^k \to x^*$. □

COROLLARY 4.2. *Let* $f \in C^1$ *be nonnegative, coercive, strictly convex, and such that the set* (iv) $\{X : f(X) \le f(X^0)\}$ *is compact. Then the sequence* $\{Z^k, X^k\}$ *generated by algorithm* (4.5)–(4.6) *converges toward* $X^* \in \Gamma$.

Proof. It is known (see again Proposition 3.4, p. 214 in [4]) that with $B = P_C(\cdot)$, assumption (i) in Lemma 4.1 is fulfilled. Note in particular that for the descent property $f(Z^{k+1}) < f(X^k)$ to hold, it is needed that $X^k \in C$. Also conditions (4.8) can always be satisfied, e.g., by taking $\alpha_k = 1$. Hence assumption (iii) holds and the result follows by applying Lemma 4.1. □

REMARK 4.1. *If* $\operatorname{rank}(A) = n$ *in CPP then the assumptions in Lemma* 4.1 *are fulfilled. Note also that the restart procedure* (4.9) *does not destroy the convergence.*

In fact our experience shows that (4.9) *gives a more robust version, i.e., less sensitive to the choice of* α_{max}. *The same observation is also made by Luenberger for the unconstrained case* [22, *Chapter* 8.7].

The following example, with $m = 5$, $n = q = 4$, was used in all but one of our experiments:

$$A = \begin{bmatrix} d & 10 & 71 & 58 \\ 40 & d & 2 & 24 \\ 66 & 56 & d & 94 \\ 97 & 92 & 14 & d \\ 4 & 14 & 5 & 73 \end{bmatrix}, \qquad B = \begin{bmatrix} 42 & 62 & 23 & 71 \\ 26 & 55 & 82 & 24 \\ 24 & 36 & 23 & 69 \\ 34 & 49 & 86 & 11 \\ 12 & 4 & 6 & 26 \end{bmatrix}.$$

Here d is a parameter that was varied to obtain different conditioning of the problem. The condition number $\kappa(A^TA)$ is 2, 83, 652, and 36,724 for d= 1000, 10, 30, and 40, respectively. The "solutions" for $d = 30$, X_1, X_2, X_3 of CPP with $C = C_1$, C_2, C_3, respectively, and the unconstrained, symmetric solution X_{uc} (of equations (3.8)) are listed in Table 4.1. These solutions were obtained by first iterating until machine precision and then rounding the results to four decimals, respectively. This corresponds roughly to stopping the gradient projection method at iteration 70; see Table 4.2. The matrix X_3 contains two eigenvalues equal to zero (within rounding errors). For the solution displayed in Table 4.1 the corresponding value for, e.g., $err2$ (see below) is $O(10^{-4})$. The Sylvester equations (3.8) were solved, in a standard way, by making a coordinate transformation into the eigenspace of A^TA.

All tests were performed in double precision arithmetic using MATLAB on a Sun Workstation. As starting matrix X^0 we always picked the projection of the solution of the unconstrained problem and for method (4.5)–(4.6) X^1 was obtained by one step of (4.1). We have experimented with the following three error measures:

$$err1 = \text{trace}(S^k \cdot (X^k)^T)/(||A||_F \cdot ||B||_F),$$

$$err2 = \lambda_{min}(S^k)/(\lambda_{max}(S^k) - \lambda_{min}(S^k)),$$

$$err3 = \begin{cases} ||X^k - X^{k+1}|| & \text{for methods (4.1) and (4.3),} \\ ||X^k - Z^{k+1}|| & \text{for method (4.5)–(4.6).} \end{cases}$$

Both $err1$ and $err2$ are invariant under scaling with a constant. Note that $err3 = ||X^k - P_C(X^k - \gamma\nabla f(X^k))||$ both for method (4.5)–(4.6) and the gradient projection method. This measure was suggested in [23], using $\gamma = 1$. It was shown that for C a polyhedral set, $X^k \in C$ and X^k sufficiently close to the solution, the distance from X^k to the solution is of the order $err3$.

In Table 4.3 we display the behavior of the error measures using the gradient projection method and $C = C_3$. Note that the starting matrix has the property that $\text{trace}(S^0X^0)$ is almost zero, but S^0 is far from being positive definite. Hence it is important to use $err1$ and $err2$ simultaneously. In our experience $err3$ was the most robust measure. Of course the set C_3 is not polyhedral so for this case the theory in [23] does not apply.

In Table 4.2 we list the behavior of the gradient projection method when applied to $C = C_1, C_2, C_3$, respectively. In Table 4.4 we compare the gradient projection method with the Han–Lou scheme and in Figure 4.1 we compare method (4.5)–(4.6) with the gradient projection method for $C = C_3$.

TABLE 4.1
The solutions for $d = 30$ and $C = C_1, C_2, C_3$, respectively.

$$X_1 = \begin{bmatrix} 0.2794 & 0.4903 & 0.8509 & 0 \\ 0 & 0 & 0 & 0 \\ 0.3910 & 0.5786 & 0 & 0.6560 \\ 0.0416 & 0 & 0 & 0.4372 \end{bmatrix},$$

$$X_2 = \begin{bmatrix} 0 & 0.3542 & 0.4064 & 0.0240 \\ 0.3542 & 0.1429 & 0.5253 & 0 \\ 0.4064 & 0.5253 & 0 & 0.0033 \\ 0.0240 & 0 & 0.0033 & 0.6888 \end{bmatrix},$$

$$X_3 = \begin{bmatrix} 0.1463 & 0.2378 & 0.2469 & 0.0654 \\ 0.2378 & 0.4315 & 0.4356 & -0.0693 \\ 0.2469 & 0.4356 & 0.4429 & -0.0234 \\ 0.0654 & -0.0693 & -0.0234 & 0.7113 \end{bmatrix},$$

$$X_{uc} = \begin{bmatrix} -1.0161 & 1.4177 & 0.7538 & -0.0430 \\ 1.4177 & -0.9716 & 0.1694 & 0.0523 \\ 0.7538 & 0.1694 & -0.2879 & 0.0841 \\ -0.0430 & 0.0523 & 0.0841 & 0.6700 \end{bmatrix}.$$

TABLE 4.2
Convergence behavior of the gradient projection method.

		$err3$		
d	k	$C = C_1$	$C = C_2$	$C = C_3$
1000	0	$1 \cdot 10^{-16}$	$1 \cdot 10^{-16}$	$5 \cdot 10^{-4}$
1000	5	$1 \cdot 10^{-16}$	$1 \cdot 10^{-16}$	$1 \cdot 10^{-7}$
1000	10	$1 \cdot 10^{-16}$	$1 \cdot 10^{-16}$	$4 \cdot 10^{-11}$
1000	20	$1 \cdot 10^{-16}$	$1 \cdot 10^{-16}$	$2 \cdot 10^{-16}$
10	0	0.74	0.11	0.10
10	50	$2 \cdot 10^{-4}$	$2 \cdot 10^{-4}$	$5 \cdot 10^{-4}$
10	100	$2 \cdot 10^{-6}$	$9 \cdot 10^{-6}$	$5 \cdot 10^{-6}$
10	200	$2 \cdot 10^{-10}$	$4 \cdot 10^{-8}$	$6 \cdot 10^{-10}$
30	0	1.2	0.37	0.17
30	100	$9 \cdot 10^{-5}$	$2 \cdot 10^{-4}$	$4 \cdot 10^{-4}$
30	200	$3 \cdot 10^{-7}$	$8 \cdot 10^{-7}$	$5 \cdot 10^{-6}$
30	400	$2 \cdot 10^{-12}$	$3 \cdot 10^{-11}$	$9 \cdot 10^{-10}$

TABLE 4.3
Comparison of three error measures using the gradient projection method with $C = C_3$ and $d = 30$.

k	$err1$	$err2$	$err3$
0	$2 \cdot 10^{-15}$	–0.32	0.17
50	$3 \cdot 10^{-3}$	$-7 \cdot 10^{-4}$	$4 \cdot 10^{-3}$
100	$3 \cdot 10^{-4}$	$-5 \cdot 10^{-6}$	$4 \cdot 10^{-4}$
200	$4 \cdot 10^{-6}$	$8 \cdot 10^{-10}$	$5 \cdot 10^{-6}$
400	$6 \cdot 10^{-10}$	$2 \cdot 10^{-13}$	$9 \cdot 10^{-10}$
600	$9 \cdot 10^{-14}$	$3 \cdot 10^{-16}$	$2 \cdot 10^{-13}$
800	$2 \cdot 10^{-15}$	$-2 \cdot 10^{-14}$	$2 \cdot 10^{-15}$

TABLE 4.4
Comparison of the Han–Lou method and the gradient projection method with $C = C_3$, $\alpha = r/\lambda_{min}(A^TA)$.

		$err3$		
d	k	r	H/L	GP
1000	5	0.5	$4 \cdot 10^{-7}$	$1 \cdot 10^{-7}$
1000	10	0.5	$6 \cdot 10^{-12}$	$4 \cdot 10^{-11}$
1000	20	0.5	$8 \cdot 10^{-17}$	$2 \cdot 10^{-16}$
10	200	0.25	$9 \cdot 10^{-3}$	$6 \cdot 10^{-10}$
10	200	0.26	$2 \cdot 10^{-9}$	
10	200	0.27	$3 \cdot 10^{-10}$	
10	200	0.28	$5 \cdot 10^{-10}$	
10	200	0.5	$7 \cdot 10^{-7}$	
30	200	0.23	$2 \cdot 10^{-4}$	$5 \cdot 10^{-6}$
30	400	0.23	$4 \cdot 10^{-6}$	$9 \cdot 10^{-10}$
30	800	0.23	$1 \cdot 10^{-9}$	$2 \cdot 10^{-15}$

As a general observation we find that the methods are quite sensitive to the conditioning of matrix A. For ill conditioned or even moderately ill conditioned matrices the rate of convergence of the three methods becomes slow. From Table 4.2 we see that there is little difference in the rate of convergence for different sets C_i. Also from Table 4.4 we find that for well-conditioned problems the Han–Lou and the gradient projection method behave quite similarly. The Han–Lou method was, however, more sensitive to the conditioning of the problem. The parameter α was taken $\alpha = r/\lambda_{min}(A^TA)$, where r was chosen experimentally. We found the Han–Lou scheme quite sensitive to the proper choice of steplength; cf. the case $d = 10$ in Table 4.4. A possible remedy would be to apply the version where α is adjusted iteratively; see section 6 in [17]. The fastest convergence with the gradient projection method was obtained with the steplength choice (4.2).

The behavior of method (4.5)–(4.6) versus the gradient projection method is illustrated in Figure 4.1. Here formula (4.2) was used for γ. Further, $\alpha_{max} = 10$ for $d = 1000, 10$ and $\alpha_{max} = 100, 1000$ for $d = 30, 40$, respectively. These values were obtained from numerical experiments. We remark, however, that the performance of method (4.5)–(4.6) was quite robust with respect to the chosen value of α_{max}. If, however, the restart rule (4.9) was left out, the method became more sensitive to this choice. We find that method (4.5)–(4.6) converges initially faster than the gradient projection method (which is especially noticeable for more ill conditioned problems). However, unlike the behavior for the gradient projection method the error does not decrease monotonically. For large iteration numbers (i.e., close to the solution) $|\bar{\alpha}_k|$ becomes large, since it corresponds to an unconstrained linesearch. Hence by the steplength rule $\alpha_k = 1$, and thus method (4.5)–(4.6) gets the same rate of convergence (although starting from a lower error level) as the gradient projection method; cf. the cases $d = 10, 30, 40$ in Figure 4.1.

Allwright [1] developed an algorithm for $C = C_3$ using a canonical hull characterization of C_3. He applied the method on a 3×3 example. In Table 4.5 we have compared the behavior of the gradient projection method and method (4.5)–(4.6) (using $\alpha_{max} = 10$) for Allwright's example. In [1] it is reported that an accuracy of order 10^{-9} is obtained after about 20 iterations. It is hard to compare these results with those in Table 4.5 since the measure of accuracy is different from ours and the complexity of each iteration is also different.

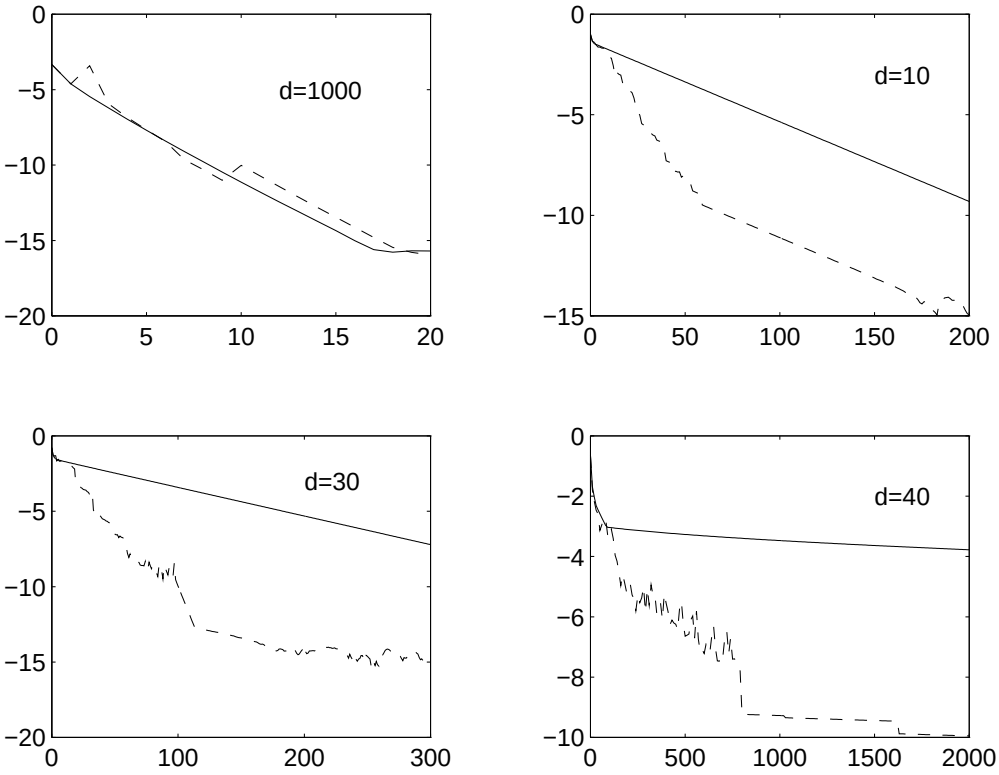


FIG. 4.1. *Error* (log(*err*3)) *versus iteration number for the gradient projection method and method* (4.5)–(4.6) *(broken curve).*

TABLE 4.5
Comparison of method (4.5)–(4.6) *and the gradient projection method using Allwright's example.*

	*err*3	
k	(4.5) − (4.6)	GP
0	0.17	0.17
25	$4 \cdot 10^{-6}$	$3 \cdot 10^{-3}$
50	$2 \cdot 10^{-8}$	$2 \cdot 10^{-4}$
75	$9 \cdot 10^{-11}$	$1 \cdot 10^{-5}$
100	$5 \cdot 10^{-12}$	$6 \cdot 10^{-7}$
125	$3 \cdot 10^{-13}$	$3 \cdot 10^{-8}$
150	$2 \cdot 10^{-14}$	$2 \cdot 10^{-9}$
200	$3 \cdot 10^{-16}$	$6 \cdot 10^{-12}$

Obviously there is a need to further study the algorithmic side of CPP in order to find a reasonable robust and fast algorithm. We leave this for future research.

Acknowledgments. We wish to thank Dr. N. J. Higham for informing us about the works of Woodgate [29] and Allwright [1]. The constructive criticism from the referees was also helpful.

REFERENCES

[1] J. C. Allwright, *Positive semidefinite matrices: Characterization via conical hulls and least-squares solution of a matrix equation,* SIAM J. Control Optim., 26 (1988), pp. 537–556.

[2] C. A. Beattie and S. W. Smith, *Optimal matrix approximants in structural identification,* J. Optim. Theory Appl., 74 (1992), pp. 23–56.

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

[4] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computing,* Prentice–Hall, Englewood Cliffs, NJ, 1989.

[5] Å. Björck, *A direct method for sparse least squares problems with lower and upper bounds*, Numer. Math., 54 (1988), pp. 19–32.

[6] Å. Björck, *Numerical Methods for Least Squares Problems,* Society for Industrial and Applied Mathematics, Philadelphia, PA, 1996.

[7] J. E. Brock, *Optimal matrices describing linear systems*, AIAA J., 6 (1968), pp. 1292–1296.

[8] 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.

[9] K. Fan and A. J. Hoffman, *Some matrix inequalities in the space of matrices,* Proc. Amer. Math. Soc., 6 (1955), pp. 111–116.

[10] R. Fletcher, *Practical Methods of Optimization,* 2nd edition, Wiley, New York, 1987.

[11] A. Friedlander and J. M. Martinez, *On the maximization of a concave quadratic function with boxconstraints,* SIAM J. Optim., 4 (1994), pp. 177–192.

[12] M. S. Gowda, *Minimising quadratic functionals over closed convex cones,* Bull. Austral. Math. Soc., 39 (1989), pp. 15–20.

[13] G. H. Golub and C. F. Van Loan, *Matrix Computations,* 2nd edition, The Johns Hopkins University Press, London, 1989.

[14] M. Hall Jr., *Combinatorial Theory,* 2nd edition, Wiley, New York, 1986.

[15] P. Halmos, *Positive approximants of operators,* Indiana Univ. Math. J., 21 (1972), pp. 951–960.

[16] S.-P. Han, *A successive projection method,* Math. Programming, 40 (1988), pp. 1–14.

[17] S.-P. Han and G. Lou, *A parallel algorithm for a class of convex programs,* SIAM J. Control Optim., 26 (1988), pp. 345–355.

[18] N. J. Higham, *The symmetric Procrustes problem,* BIT, 28 (1988), pp. 133–143.

[19] N. J. Higham, *Computing a nearest symmetric positive semidefinite matrix,* Linear Algebra Appl., 103 (1988), pp. 103–118.

[20] E. Isaacson and H. B. Keller, *Analysis of Numerical Methods,* Wiley, New York, 1966.

[21] A. N. Iusem and A. R. De Pierro, *On the convergence of Han's method for convex programming with quadratic objective,* Math. Programming, 52 (1991), pp. 265–284.

[22] D. G. Luenberger, *Linear and Nonlinear Programming,* 2nd ed., Addison–Wesley, Reading, MA, 1984.

[23] Z.-Q. Luo and P. Tseng, *On the linear convergence of descent methods for convex essentially smooth minimization,* SIAM J. Control Optim., 30 (1992), pp. 408–425.

[24] R. T. Rockafellar, *Convex analysis,* Princeton University Press, Princeton, NJ, 1970.

[25] A. Ruhe, *Closest normal matrix finally found!,* BIT, 27 (1987), pp. 585–598.

[26] J. Stoer and C. Witzgall, *Convexity and Optimization in Finite Dimensions* I, Springer-Verlag, New York, 1970.

[27] G. Strang, *Introduction to Applied Mathematics,* Wellesley–Cambridge Press, Wellesley, MA, 1986.

[28] T. J. Suffridge and T. L. Hayden, *Approximation by a Hermitian positive semidefinite Toeplitz matrix,* SIAM J. Matrix Anal. Appl., 14 (1993), pp. 721–734.

[29] K. G. Woodgate, *Least-Squares Solution of $F = PG$ over Positive Semi-Definite Symmetric P^**, Technical Report, Dept. of Aeronautics, Imperial College of Science, Technology and Medicine, London, January 1993.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 140–158, January 1997

012

AN UNSYMMETRIC-PATTERN MULTIFRONTAL METHOD FOR SPARSE LU FACTORIZATION*

TIMOTHY A. DAVIS† AND IAIN S. DUFF‡

Abstract. Sparse matrix factorization algorithms for general problems are typically characterized by irregular memory access patterns that limit their performance on parallel-vector supercomputers. For symmetric problems, methods such as the multifrontal method avoid indirect addressing in the innermost loops by using dense matrix kernels. However, no efficient LU factorization algorithm based primarily on dense matrix kernels exists for matrices whose pattern is very unsymmetric. We address this deficiency and present a new unsymmetric-pattern multifrontal method based on dense matrix kernels. As in the classical multifrontal method, advantage is taken of repetitive structure in the matrix by factorizing more than one pivot in each frontal matrix, thus enabling the use of Level 2 and Level 3 BLAS. The performance is compared with the classical multifrontal method and other unsymmetric solvers on a CRAY C-98.

Key words. LU factorization, unsymmetric sparse matrices, multifrontal methods

AMS subject classifications. 65F50, 65F05, 65Y15, 65-04

PII. S089547989324690X

NOTATION.

$\mathbf{A}$	original matrix
$\mathbf{A}^k$	undeleted rows and columns of the original matrix at step k
$\mathcal{A}$	Struct($\mathbf{A}$)
$\mathbf{A}'$	active submatrix
$\mathbf{F}$	current frontal matrix
$\mathbf{C}$	contribution block of $\mathbf{F}$
$\mathbf{L}'$, $\mathbf{L}''$	the $\lvert\mathcal{L}'\rvert$ columns of $\mathbf{L}$ computed in $\mathbf{F}$
$\widehat{\mathbf{L}}$	the portion of $\mathbf{L}''$ whose updates have yet to be applied to $\mathbf{C}$
$\mathbf{U}'$, $\mathbf{U}''$	the $\lvert\mathcal{U}'\rvert$ rows of $\mathbf{U}$ computed in $\mathbf{F}$
$\widehat{\mathbf{U}}$	the portion of $\mathbf{U}''$ whose updates have yet to be applied to $\mathbf{C}$
$\mathcal{L}$	sequence of row indices of $\mathbf{F}$ (union of pattern of pivotal columns in $\mathbf{F}$)
$\mathcal{L}'$	pivotal row indices in $\mathcal{L}$
$\mathcal{L}''$	nonpivotal row indices in $\mathcal{L}$ ($\mathcal{L} = \mathcal{L}' \cup \mathcal{L}''$)
$\mathcal{U}$	sequence of column indices of $\mathbf{F}$ (union of pattern of pivotal rows in $\mathbf{F}$)
$\mathcal{U}'$	pivotal column indices in $\mathcal{U}$
$\mathcal{U}''$	nonpivotal column indices in $\mathcal{U}$ ($\mathcal{U} = \mathcal{U}' \cup \mathcal{U}''$)
$\overline{V}$	a list of the first pivotal indices of the factorized frontal matrices
e	an element $e \in \overline{V}$

* Received by the editors November 15, 1995; accepted for publication (in revised form) by J. W. H. Liu February 3, 1996.
http://www.siam.org/journals/simax/18-1/24690.html

† Computer and Information Science and Engineering Department, University of Florida, Gainesville, FL 32611-6120 (davis@cise.ufl.edu). Technical reports, software, and matrices are available via the World Wide Web at http://www.cise.ufl.edu/~davis or by anonymous ftp at ftp.cise.ufl.edu. Support for this project was provided by National Science Foundation grants ASC-9111263 and DMS-9223088 and by CRAY Research, Inc. and Florida State University through the allocation of supercomputer resources. Portions of this work were supported by a postdoctoral grant from CERFACS.

‡ Rutherford Appleton Laboratory, Chilton, Didcot, Oxon 0X11 0QX, England, and European Center for Research and Advanced Training in Scientific Computation (CERFACS), Toulouse, France (isd@letterbox.rl.ac.uk). Technical reports, information on the Harwell Subroutine Library, and matrices are available via the World Wide Web at http://www.cis.rl.ac.uk/struct/ARCD/NUM.html or by anonymous ftp at seamus.cc.rl.ac.uk/pub.

$\mathbf{C}_e$	remaining portions of a previous contribution block ($e < k$)
$\mathcal{L}_e$	row indices of $\mathbf{C}_e$
$\mathcal{U}_e$	column indices of $\mathbf{C}_e$
$\mathcal{C}_j$	element list of column j of $\mathbf{A}'$
$\mathcal{R}_i$	element list of row i of $\mathbf{A}'$
$d_r(i)$	the true degree of row i
$\overline{d_r}(i)$	upper bound of the degree of row i
$d_c(j)$	the true degree of column j
$\overline{d_c}(j)$	upper bound of the degree of column j
$w()$	a work array for computing external column degrees
$\|\ldots\|$	number of entries in a matrix or set, or absolute value of a scalar, depending on the context
Struct(...)	row indices of entries in a column, or column indices of entries in a row

1. Introduction. Conventional sparse matrix factorization algorithms for general problems rely heavily on indirect addressing. This gives them an irregular memory access pattern that limits their performance on typical parallel-vector supercomputers and on cache-based RISC architectures. In contrast, the multifrontal method of Duff [8], Duff, Erisman, and Reid [9], and Duff and Reid [13, 14] is designed with regular memory access in the innermost loops and has been modified by Amestoy and Duff to use standard kernels [2]. This multifrontal method assumes structural symmetry and bases the factorization on an *assembly tree* generated from the original matrix and an ordering such as minimum degree. The computational kernel, executed at each node of the tree, is one or more steps of LU factorization within a square, dense *frontal matrix* defined by the nonzero pattern of a pivot row and column. These steps of LU factorization compute a *contribution block* (a Schur complement) that is later *assembled* (added) into the frontal matrix of its parent in the assembly tree. Henceforth we will call this approach the *classical* multifrontal method.

Although structural asymmetry can be accommodated in the classical multifrontal method by holding the pattern of $\mathbf{A}+\mathbf{A}^T$ and storing explicit zeros, this can have poor performance on matrices whose patterns are very unsymmetric. If we assume from the outset that the matrix may be structurally asymmetric, the situation becomes more complicated. For example, the frontal matrices are rectangular instead of square, and some contribution blocks must be assembled into more than one subsequent frontal matrix. As a consequence, it is no longer possible to represent the factorization by an assembly tree and the more general structure of an *assembly dag* (directed acyclic graph) [5] similar to that of Gilbert and Liu [20] and Eisenstat and Liu [16, 17, 18] is required. In the current work we do not explicitly use this structure. Since we consider an algorithm that combines the symbolic analysis and numerical factorization, our algorithm for a subsequent numerical factorization (which uses a dag) is beyond the scope of this paper.

We have developed a new unsymmetric-pattern multifrontal approach [4, 5]. As in the symmetric multifrontal case, advantage is taken of repetitive structure in the matrix by factorizing more than one pivot in each frontal matrix. Thus the algorithm can use higher level dense matrix kernels in its innermost loops (Level 3 BLAS [6]). We refer to the unsymmetric-pattern multifrontal method described in this paper as UMFPACK Version 1.0 [4]. A parallel factorize-only version of UMFPACK, based on the assembly dag, is discussed in Hadfield's dissertation [23] and related work [24, 25]. The multifrontal method for symmetric positive definite matrices is reviewed in [27].

Section 2 presents an overview of the basic approach and a brief outline of

the algorithm. We introduce our data structures in the context of a small sparse matrix in section 3, where we describe the factorization of the first frontal matrix. In section 4 we develop the algorithm further by discussing how subsequent frontal matrices are factorized. We have split the discussion of the algorithm into these two sections so that we can define important terms in the earlier section while considering a less complicated situation. Section 5 presents a full outline of the algorithm, using the notation introduced in previous sections. In section 6, we compare the performance of our algorithm with an algorithm based on the classical multifrontal method (MUPS, [2]), and an algorithm based on conventional (compressed sparse vector) data structures (MA48, [15]).

2. The basic approach. Our goal with the UMFPACK algorithm is to achieve high performance in a general unsymmetric sparse factorization code by using the Level 3 BLAS. We accomplish this by developing a multifrontal technique that uses rectangular frontal matrices and chooses several pivots within each frontal matrix. High performance is also achieved through an approximate degree update algorithm that is much faster (asymptotically and in practice) than computing the true degrees. A general sparse code must select pivots based on both numerical and symbolic (fill-reducing) criteria. We therefore combine the analysis phase (pivot selection and symbolic factorization) with the numerical factorization. We construct our rectangular frontal matrices dynamically, since we do not know their structure prior to factorization. An assembly dag is constructed during this analyze–factorize phase. We use the assembly dag in the factorize-only phase, and Hadfield [23] and Hadfield and Davis [24, 25] develop it further and use it in a parallel factorize-only algorithm.

The *active matrix* is the Schur complement of $\mathbf{A}$ that remains to be factorized. At a particular stage, the frontal matrix is initialized through choosing a pivot from anywhere in the active matrix (called a global pivot search) using a Zlatev-style pivot search [29], except that we keep track of upper bounds on the degrees of rows and columns in the active matrix, rather than the true degrees. (The degree of a row or column is simply the number of entries in the row or column.) We call this first pivot the *seed* pivot. Storage for the frontal matrix is allocated to contain the entries in the pivot row and column plus some room for further expansion determined by an input parameter. We denote the current frontal matrix by $\mathbf{F}$ and the submatrix comprising the rows and columns not already pivotal by $\mathbf{C}$, calling $\mathbf{C}$ the contribution block.

Subsequent pivots within this frontal matrix are found within the contribution block $\mathbf{C}$, as shown in Figure 2.1. The frontal matrix grows as more pivots are chosen, as denoted by the arrows in the figure. We assemble contribution blocks from earlier frontal matrices into this frontal matrix as needed. The selection of pivots within this frontal matrix stops when our next choice for pivot would cause the frontal matrix to become larger than the allocated working array. We then complete the factorization of the frontal matrix using Level 3 BLAS, store the LU factors, and place the contribution block $\mathbf{C}$ in a heap. The contribution block is deallocated when it is assembled into a subsequent frontal matrix. We then continue the factorization by choosing another seed pivot and generating and factorizing a new frontal matrix.

It is too expensive to compute the actual degrees of the rows and columns of the active matrix. To do so would require at least as much work as the numerical factorization itself. This would defeat the performance gained from using the dense matrix kernels. Instead, we compute upper bounds for these degrees at a much lower complexity than the true degrees, since they are obtained from the frontal matrix data structures instead of conventional sparse vectors. We avoid forming the union of

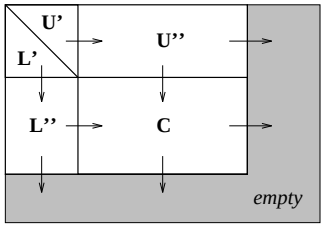


FIG. 2.1. *A rectangular frontal matrix within a larger working array.*

sparse rows or columns which would have been needed were we to compute the filled patterns of rows and columns in the active matrix. We have incorporated a symmetric analogue of our approximate degree update algorithm into the approximate minimum degree (AMD) ordering algorithm [1]. The algorithm produces the same quality of ordering as prior minimum degree ordering algorithms and is typically faster.

The performance we achieve in the UMFPACK algorithm thus depends equally on two crucial factors: this approximate degree update algorithm and the numerical factorization within dense, rectangular frontal matrices. An outline of the UMFPACK algorithm is shown in Algorithm 1. If $\mathbf{A}$ is permuted to block upper triangular form [11], the algorithm is applied to each block on the diagonal. Algorithm 1 consists of initializations followed by three steps, as follows:

ALGORITHM 1 (outline of the unsymmetric-pattern multifrontal algorithm).

```
0:   initializations
     while (factorizing A) do
1:       global pivot search for seed pivot
         form frontal matrix F
         while (pivots found within frontal matrix) do
2:           assemble prior contribution blocks and original rows into F
             compute the degrees of rows and columns in C (the contribution
             block of F)
             numerically update part of C (Level 2 and Level 3 BLAS)
             local pivot search within C
         endwhile
3:       complete the factorization of F using Level 3 BLAS
         place C in heap
     endwhile
```

The initialization phase of the algorithm (step 0) converts the original matrix into two compressed sparse vector forms (row oriented and column oriented [9]) with numerical values $\mathbf{A}$ and symbolic pattern $\mathcal{A}$. Rows and columns are used and deleted from $\mathbf{A}$ and $\mathcal{A}$ during factorization when they are assembled into frontal matrices. At any given step, k say, we use $\mathbf{A}^k$ and $\mathcal{A}^k$ to refer to entries in the original matrix that are not yet deleted. An *entry* is defined by a value in the matrix that is actually stored. Thus all nonzeros are entries but some entries may have the value zero. We use $|\ldots|$ both to denote the absolute value of a scalar and to signify the number of entries in a set, sequence, or matrix. The meaning should always be quite clear from the context.

The true degrees $d_r(i)$ and $d_c(j)$ are the number of entries in row i and column j of the active matrix $\mathbf{A}'$, respectively, but we do not store these. Because the cost

of updating these would be prohibitive, we instead use upper bounds $\overline{d_r}(i)$ (where $d_r(i) \leq \overline{d_r}(i)$) and $\overline{d_c}(j)$ (where $d_c(j) \leq \overline{d_c}(j)$). However, when a true degree is computed, as in the initialization phase or during the search for a seed pivot, its corresponding upper bound is set equal to the true degree.

3. The first frontal matrix. We will label the frontal matrix generated at stage e by the index e. We now describe the factorization of the first frontal matrix ($e = 1$). This discussion is, however, also applicable for subsequent frontal matrices ($e > 1$) which are discussed in full in section 4 where differences from the case $e = 1$ are detailed.

3.1. Step 1: Perform global pivot search and form frontal matrix. The algorithm performs pivoting both to maintain numerical stability and to reduce fill in. The first pivot in each frontal matrix is chosen using a global Zlatev-style search [29]. A few candidate columns with the lowest upper bound degrees are searched. The number searched is controlled by an input parameter (which we denote by *nsrch* and whose default value is four). Among those *nsrch* columns, we select as pivot the entry a'_{rc} with the smallest approximate Markowitz cost [28], $(\overline{d_r}(r) - 1)(d_c(c) - 1)$, such that a'_{rc} also satisfies a threshold partial pivoting condition [9]

$$|a'_{rc}| \geq u \cdot \max_i |a'_{ic}|, \quad 0 < u \leq 1. \tag{3.1}$$

Note that we have the true column degree. The column entries are just the entries in $\mathbf{A}$ since this is the first frontal matrix. When the pivot is chosen its row and column structure define the frontal matrix. If Struct(...) denotes the row indices of entries in a column or column indices of entries in a row, we define $\mathcal{L}$ and $\mathcal{U}$ by $\mathcal{L} = \text{Struct}(\mathbf{A}'_{*c})$ and $\mathcal{U} = \text{Struct}(\mathbf{A}'_{r*})$, the row and column indices, respectively, of the current $|\mathcal{L}|$-by-$|\mathcal{U}|$ frontal matrix $\mathbf{F}$. We partition the sets $\mathcal{L}$ and $\mathcal{U}$ into pivotal row and column indices ($\mathcal{L}'$ and $\mathcal{U}'$) and nonpivotal row and column indices ($\mathcal{L}''$ and $\mathcal{U}''$).

We then assemble the pivot row ($\mathbf{A}^k_{r*}$) and column ($\mathbf{A}^k_{*c}$) from the original matrix into $\mathbf{F}$ and delete them from $\mathbf{A}^k$ (which also deletes them from $\mathcal{A}^k$, since $\mathcal{A}^k$ is defined as Struct($\mathbf{A}^k$)).

We then try to find further pivot rows and columns with identical pattern in the same frontal matrix. This process is called *amalgamation*. *Relaxed amalgamation* does the same with pivots of similar but nonidentical pattern. To permit relaxed amalgamation, $\mathbf{F}$ is placed in the upper left corner of a larger, newly allocated, s-by-t work array. Relaxed amalgamation is controlled by choosing values for s and t through the input parameter g, where $s = \lfloor g|\mathcal{L}| \rfloor$, $t = \lfloor g|\mathcal{U}| \rfloor$, and $g \geq 1$. The default value of this parameter in UMFPACK is $g = 2$.

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 & 0 & a_{14} & a_{15} & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 & a_{25} & 0 & 0 \\ a_{31} & a_{32} & a_{33} & 0 & 0 & 0 & a_{37} \\ a_{41} & 0 & 0 & a_{44} & a_{45} & a_{46} & 0 \\ 0 & a_{52} & a_{53} & 0 & a_{55} & a_{56} & 0 \\ 0 & 0 & 0 & 0 & 0 & a_{66} & a_{67} \\ a_{71} & a_{72} & 0 & 0 & a_{75} & 0 & a_{77} \end{bmatrix}. \tag{3.2}$$

We use example (3.2) to illustrate our discussion in this section and in section 4. Permutations would needlessly obscure the example, so we assume the pivots in the example matrix are on the diagonal, in order. (Note that this assumption would not

TABLE 3.1
True degrees and degree bounds in example matrix.

i	$d_r(i)$	$\overline{d_r}(i)$	j	$d_c(j)$	$\overline{d_c}(j)$
2	4	5	2	4	4
3	5	5	3	3	3
4	3	3	4	4	4
5	4	4	5	5	6
6	2	2	6	3	3
7	4	5	7	3	3

be true if we performed a global pivot search as in Step 1 since in our example the pivots do not have the lowest possible Markowitz cost.) The first pivot is a'_{11}. We have $\mathcal{L} = \mathcal{L}' \cup \mathcal{L}'' = \{1,2,3,4,7\} = \{1\} \cup \{2,3,4,7\}$ and $\mathcal{U} = \mathcal{U}' \cup \mathcal{U}'' = \{1,4,5\} = \{1\} \cup \{4,5\}$. Let g be 1.25; then the 5-by-3 frontal matrix would be stored in a 6-by-3 array.

3.2. Step 2: Choose further pivots, perform assemblies, and partial factorization. We continue our pivot search within the contribution block $\mathbf{C}$ of the current frontal matrix $\mathbf{F}$ and repeat this for as long as there is sufficient space in the working array.

We use the term *assembly* for the addition of contribution terms or original entries via the extend-add ("$\boxplus$") operator [27]. This operator aligns the row and column index sets of its two matrix or vector operands and then adds together values referenced by the same indices. An *implicit* assembly is one that is mathematically represented by the data structures but computationally postponed. An *explicit* assembly is one that is actually computed. An entry in the active matrix $\mathbf{A}'$ is explicitly assembled if all its contribution terms have been added to it, but this is usually not done and such entries are normally only held implicitly. Pivotal rows and columns are always explicitly assembled.

We now describe the test to determine whether a column can be assembled into $\mathbf{F}$. We scan $\mathcal{A}^k_{*j}$ for each column j in $\mathcal{U}''$. The scan of $\mathcal{A}^k_{*j}$ is stopped as soon as a row $i \notin \mathcal{L}$ is found. If the scan completes without such a row being found, then all row indices in $\mathcal{A}^k_{*j}$ are also in $\mathcal{L}$, and we delete $\mathbf{A}^k_{*j}$ from $\mathbf{A}$ and assemble it into $\mathbf{F}$. If this assembly is done, the true degree of column j is $d_c(j) = \overline{d_c}(j) = |\mathcal{L}''|$. If the scan stops early, we compute the upper bound degree of column j as

$$\overline{d_c}(j) = \min \left\{ \begin{array}{ll} n-k & \text{(the size of } \mathbf{A}'\text{)} \\ |\mathcal{L}''| + (|\mathcal{A}^k_{*j}| - \alpha_j) & \text{(the worst case fill-in)} \end{array} \right\},$$

where k is the current step of Gaussian elimination and α_j is the number of entries scanned in $\mathcal{A}^k_{*j}$ before stopping. For each row i in $\mathcal{L}''$, we scan $\mathcal{A}^k_{i*}$ and compute $\overline{d_r}(i)$ in an analogous manner, where we define β_i as the number of entries scanned in $\mathcal{A}^k_{i*}$ before stopping.

In the example, $\mathbf{A}^k_{*4}$ is assembled into $\mathbf{C}$ and entry a_{44} is deleted from $\mathbf{A}$. The uncomputed true degrees and the degree bounds are shown in Table 3.1. The values of α_j used in constructing the upper bounds were obtained on the assumption that the rows and columns of $\mathcal{A}^k$ are stored in ascending order of row and column indices. We make this assumption only to simplify the example. We have

$$\mathbf{F} = \left\{ \begin{array}{c|cc} & \mathcal{U}' & \mathcal{U}'' \\ \hline \mathcal{L}' & a'_{rc} & \mathbf{A}'_{r*} \\ \mathcal{L}'' & \mathbf{A}'_{*c} & \mathbf{C} \end{array} \right\} = \left\{ \begin{array}{c|ccc} & 1 & 4 & 5 \\ \hline 1 & a'_{11} & a'_{14} & a'_{15} \\ 2 & a'_{21} & 0 & 0 \\ 3 & a'_{31} & 0 & 0 \\ 4 & a'_{41} & a_{44} & 0 \\ 7 & a'_{71} & 0 & 0 \end{array} \right\}.$$

We divide the pivot column $\mathbf{A}'_{*c}$ by the pivot a'_{rc} to obtain the kth column of $\mathbf{L}$, the n-by-n lower triangular factor. The pivot row is the kth row of $\mathbf{U}$, the n-by-n upper triangular factor. Step k of Gaussian elimination is complete, except for the updates from the kth pivot. The counter k is now incremented for the next step of Gaussian elimination. The frontal matrix $\mathbf{F}$ is partitioned into four submatrices, according to the partition of $\mathcal{L}$ and $\mathcal{U}$. We have

$$\mathbf{F} = \left\{ \begin{array}{c|cc} & \mathcal{U}' & \mathcal{U}'' \\ \hline \mathcal{L}' & \mathbf{L}'\mathbf{U}' & \mathbf{U}'' \\ \mathcal{L}'' & \mathbf{L}'' & \mathbf{C} \end{array} \right\} = \left\{ \begin{array}{c|ccc} & 1 & 4 & 5 \\ \hline 1 & u_{11} & u_{14} & u_{15} \\ 2 & l_{21} & 0 & 0 \\ 3 & l_{31} & 0 & 0 \\ 4 & l_{41} & a_{44} & 0 \\ 7 & l_{71} & 0 & 0 \end{array} \right\}.$$

The updates to $\mathbf{C}$ from the $|\mathcal{U}'|$ pivots in $\mathbf{F}$ are not applied one at a time. Instead, they are delayed until there are updates pending from b pivots to allow the efficient use of Level 3 BLAS [6]. On a CRAY C-98, a good value for the parameter b is 16. Let $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{U}}$ denote the portions of $\mathbf{L}''$ and $\mathbf{U}''$, respectively, whose updates have yet to be fully applied to $\mathbf{C}$. If $|\mathcal{U}'| \bmod b = 0$ then the pending updates are applied ($\mathbf{C} = \mathbf{C} - \widehat{\mathbf{L}}\widehat{\mathbf{U}}$). If b were 16, no updates would be applied in our example since $|\mathcal{U}'| = 1$.

We now search for the next pivot within the current frontal matrix. We search the columns in $\mathcal{U}''$ to find a candidate pivot column c that has minimum $\overline{d_c}(c)$ among the columns of $\mathcal{U}''$. We then apply any pending updates to this candidate column ($\mathbf{C}_{*c} = \mathbf{C}_{*c} - \widehat{\mathbf{L}}\widehat{\mathbf{U}}_{*c}$) and compute the candidate column $\mathbf{A}'_{*c}$, its pattern $\mathrm{Struct}(\mathbf{A}'_{*c})$, and its true degree $d_c(c)$. (If the updated candidate column is not selected as a pivot, it is not necessary to update it in Step 3 of the algorithm, discussed below in section 3.3.) We select the candidate pivot row r in $\mathcal{L}''$ with the lowest $\overline{d_r}(r)$ such that a'_{rc} also satisfies the threshold pivoting criterion (equation (3.1)). We compute the pattern $\mathrm{Struct}(\mathbf{A}'_{r*})$ of the candidate pivot row and its true degree $d_r(r)$.

If $d_c(c) > s - |\mathcal{U}'|$ or $d_r(r) > t - |\mathcal{U}'|$ the current work array is too small to accommodate the candidate pivot and we stop the pivot search. Also, if the candidate column has entries outside the current frontal matrix, the threshold pivoting criterion might prevent us from finding an acceptable candidate pivot in $\mathcal{L}''$. In this case also we stop the factorization of the current frontal matrix $\mathbf{F}$. If the candidate pivot a'_{rc} is acceptable, then we let $\mathcal{L} = \mathcal{L} \cup \mathrm{Struct}(\mathbf{A}'_{*c})$ and $\mathcal{U} = \mathcal{U} \cup \mathrm{Struct}(\mathbf{A}'_{r*})$. We repartition $\mathcal{L}$ and $\mathcal{U}$ into pivotal row and column indices ($\mathcal{L}'$ and $\mathcal{U}'$) and nonpivotal row and column indices ($\mathcal{L}''$ and $\mathcal{U}''$) and apply any pending updates to the pivot row ($\mathbf{C}_{r*} = \mathbf{C}_{r*} - \widehat{\mathbf{L}}_{r*}\widehat{\mathbf{U}}$).

In the example, the candidate column (column 4) can fit in the 6-by-3 work array (that is, $d_c(4) = 4 \le s - |\mathcal{U}'| = 6 - 1 = 5$). Suppose a'_{44} does not meet the threshold criterion, and row 7 is selected as the candidate row. The candidate row is, however,

rejected when its true degree is computed (the work array is too small to accommodate row 7, since $d_r(7) = 4 > t - |\mathcal{U}'| = 3 - 1 = 2$).

3.3. Step 3: Complete the factorization of F. After the last pivot has been selected within the current frontal matrix $\mathbf{F}$, we apply any pending updates to the contribution block. ($\mathbf{C} = \mathbf{C} - \widehat{\mathbf{L}}\widehat{\mathbf{U}}$, but we do not need to update the failed candidate pivot column, if any.) The pivot rows and columns in $\mathbf{F}$ are then placed in storage allocated for the LU factors.

The contribution block $\mathbf{C}$ and its pattern $\mathcal{L}''$ and $\mathcal{U}''$ form what we call an *element*. In particular, let $\mathbf{C}_e$ denote the contribution block of element e, and let the pattern of $\mathbf{C}_e$ be $\mathcal{L}_e$ and $\mathcal{U}_e$ (note that $\mathcal{L}_e = \mathcal{L}''$ and $\mathcal{U}_e = \mathcal{U}''$). The contribution block $\mathbf{C}_e$ is placed in a heap for assembly into subsequent frontal matrices.

Initially, all row and column indices in $\mathcal{L}_e$ and $\mathcal{U}_e$ are unmarked. When a row (or column) of $\mathbf{C}_e$ is assembled into a subsequent frontal matrix, the corresponding index is marked in $\mathcal{L}_e$ (or $\mathcal{U}_e$). Element e (which consists of the terms $\mathbf{C}_e$, $\mathcal{L}_e$, and $\mathcal{U}_e$) will refer to unmarked portions only. Element e is deleted when all of its entries are assembled into subsequent frontal matrices. For our example, element e is

$$\left\{ \begin{array}{c} \\ 2 \\ 3 \\ 4 \\ 7 \end{array} \begin{array}{c} \begin{array}{cc} 4 & 5 \end{array} \\ \left[\begin{array}{cc} c_{24} & c_{25} \\ c_{34} & c_{35} \\ c_{44} & c_{45} \\ c_{74} & c_{75} \end{array} \right] \end{array} \right\}.$$

We associate with each row (column) in the active matrix an *element list*, which is a list of the elements that hold pending updates to the row (column). We denote the list of elements containing row i as $\mathcal{R}_i$ and the list of elements containing column j as $\mathcal{C}_j$. The element lists contain a local index which identifies which row or column in the element matrix is equivalent to the row or column of the active matrix. This facilitates the numerical assembly of individual rows and columns. For each row i in $\mathcal{L}_e$, we place an element/local-index pair (e, m) in the element list $\mathcal{R}_i$, where row i is the mth entry of $\mathcal{L}_e$. Similarly, for each column j in $\mathcal{U}_e$, we place (e, m) in the element list $\mathcal{C}_j$, where column j is the mth entry of $\mathcal{U}_e$.

Let $\sum^{\dotplus}$ denote a summation using the $\dotplus$ operator. The active matrix $\mathbf{A}'$ is represented by an implicit assembly of $\mathbf{A}^k$ and the elements in the set $\overline{V}$,

$$\mathbf{A}' = \left(\sum_{e \in \overline{V}}^{\dotplus} \mathbf{C}_e \right) \dotplus \mathbf{A}^k, \tag{3.3}$$

where $\overline{V} \subseteq \{1 \ldots k-1\}$ is the set of elements that remain after step $k-1$ of Gaussian elimination. All $\dotplus$ operations in equation (3.3) are not explicitly performed and are postponed, unless stated otherwise. As defined earlier, the notation $\mathbf{A}^k$ refers to original entries in nonpivotal rows and columns of the original matrix that have not yet been assembled into any frontal matrices.

The element lists allow equation (3.3) to be evaluated one row or column at a time, as needed. Column j of $\mathbf{A}'$ is

$$\mathbf{A}'_{*j} = \left(\sum_{(e,m) \in \mathcal{C}_j}^{\dotplus} [\mathbf{C}_e]_{*m} \right) \dotplus \mathbf{A}^k_{*j} \tag{3.4}$$

TABLE 3.2
Element lists for example matrix after first frontal matrix.

i	$\mathcal{R}_i$	j	$\mathcal{C}_j$
2	(1,1)	2	-
3	(1,2)	3	-
4	(1,3)	4	(1,1)
5	-	5	(1,2)
6	-	6	-
7	(1,4)	7	-

with pattern

$$\text{Struct}(\mathbf{A}'_{*j}) = \left(\bigcup_{e \in \mathcal{C}_j} \mathcal{L}_e \right) \cup \mathcal{A}^k_{*j}. \tag{3.5}$$

Similarly, row i of $\mathbf{A}'$ is

$$\mathbf{A}'_{i*} = \left(\sum_{(e,m) \in \mathcal{R}_i}^{\Updownarrow} [\mathbf{C}_e]_{m*} \right) \Updownarrow \mathbf{A}^k_{i*} \tag{3.6}$$

with pattern

$$\text{Struct}(\mathbf{A}'_{i*}) = \left(\bigcup_{e \in \mathcal{R}_i} \mathcal{U}_e \right) \cup \mathcal{A}^k_{i*}. \tag{3.7}$$

There is an interesting correspondence between our data structures and George and Liu's quotient graph representation of the factorization of a symmetric positive definite matrix [19]. Suppose we factorize a symmetric positive definite matrix using our algorithm and restrict the pivots to the diagonal. Then $\mathcal{A}^k_{i*} = \mathcal{A}^k_{*i}$, $\mathcal{R}_i = \mathcal{C}_i$, $\mathcal{L}_e = \mathcal{U}_e$, and $\text{Adj}_{\mathcal{G}_k}(x_i) = \mathcal{R}_i \cup \mathcal{A}^k_{i*}$, where x_i is an uneliminated node in the quotient graph $\mathcal{G}_k$. The uneliminated node x_i corresponds to a row i and column i in $\mathbf{A}'$. That is, the sets $\mathcal{R}_i$ and $\mathcal{A}^k_{i*}$ are the eliminated supernodes and uneliminated nodes, respectively, that are adjacent to the uneliminated node x_i. In our terminology, the eliminated supernode $\overline{x}_e$ corresponds to element $e \in \overline{V}$. The set $\mathcal{L}_e$ contains the uneliminated nodes that are adjacent to the eliminated supernode $\overline{x}_e$. That is, $\text{Adj}_{\mathcal{G}_k}(\overline{x}_e) = \mathcal{L}_e$.

After the first frontal matrix on example (3.2), $\overline{V} = \{1\}$ and

$$\mathbf{A}' = \mathbf{C}_1 \Updownarrow \mathbf{A}^k = \left\{ \begin{array}{c} \\ 2 \\ 3 \\ 4 \\ 7 \end{array} \begin{array}{c} \begin{array}{cc} 4 & 5 \end{array} \\ \left[\begin{array}{cc} c_{24} & c_{25} \\ c_{34} & c_{35} \\ c_{44} & c_{45} \\ c_{74} & c_{75} \end{array} \right] \end{array} \right\} \Updownarrow \left\{ \begin{array}{c} \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccc} 2 & 3 & 5 & 6 & 7 \end{array} \\ \left[\begin{array}{ccccc} a_{22} & a_{23} & a_{25} & 0 & 0 \\ a_{32} & a_{33} & 0 & 0 & a_{37} \\ 0 & 0 & a_{45} & a_{46} & 0 \\ a_{52} & a_{53} & a_{55} & a_{56} & 0 \\ 0 & 0 & 0 & a_{66} & a_{67} \\ a_{72} & 0 & a_{75} & 0 & a_{77} \end{array} \right] \end{array} \right\}.$$

Note that column four was deleted from $\mathbf{A}^k$ (refer to section 3.2). It also no longer appears in $\mathcal{A}^k$. The element lists are given in Table 3.2. Applying

equations (3.6) and (3.7) to obtain row two, for example, we obtain

$$\mathbf{A}'_{2*} = [\mathbf{C}_1]_{1*} \Updownarrow \mathbf{A}^k_{2*} = \left\{ \begin{array}{cc} & \begin{array}{cc} 4 & 5 \end{array} \\ 2 & \left[\begin{array}{cc} c_{24} & c_{25} \end{array} \right] \end{array} \right\} \Updownarrow \left\{ \begin{array}{cc} & \begin{array}{ccc} 2 & 3 & 5 \end{array} \\ 2 & \left[\begin{array}{ccc} a_{22} & a_{23} & a_{25} \end{array} \right] \end{array} \right\}$$
$$= \left\{ \begin{array}{cc} & \begin{array}{cccc} 4 & 5 & 2 & 3 \end{array} \\ 2 & \left[\begin{array}{cccc} c_{24} & (c_{25} + a_{25}) & a_{22} & a_{23} \end{array} \right] \end{array} \right\},$$

$$\text{Struct}(\mathbf{A}'_{2*}) = \mathcal{U}_1 \cup \mathcal{A}^k_{2*} = \{4, 5\} \cup \{2, 3, 5\} = \{4, 5, 2, 3\}.$$

4. Subsequent frontal matrices. We now describe how later steps differ when the element lists are not empty by continuing the example with the second frontal matrix.

4.1. Step 1: Perform global pivot search and form frontal matrix. We compute the *nsrch* candidate pivot columns using equations (3.4) and (3.5). The assembled forms of the unused $nsrch - 1$ candidate columns are discarded. Note that this differs from how we treat an unused candidate column during the local pivot search. Updates to a single unused local candidate column are kept, as discussed in section 3.3. In the example, the next pivot is a'_{22}, with $\mathcal{L} = \mathcal{L}' \cup \mathcal{L}'' = \{2, 3, 5, 7\} = \{2\} \cup \{3, 5, 7\}$ and $\mathcal{U} = \mathcal{U}' \cup \mathcal{U}'' = \{2, 3, 4, 5\} = \{2\} \cup \{3, 4, 5\}$. The 4-by-4 frontal matrix is stored in a 5-by-5 array ($g = 1.25$).

4.2. Step 2: Choose further pivots, perform assemblies, and partial factorization. In the example, a second pivot (a'_{33}) is found in the second frontal matrix and so we will repeat this step twice.

As we discussed earlier, computing the true degree, $d_c(j) = |\text{Struct}(\mathbf{A}'_{*j})|$, with equation (3.5) would be very time consuming. A loose upper bound on $d_c(j)$ can be derived if we assume no overlap between $\mathcal{L}$ and each $\mathcal{L}_e$, viz.,

$$d_c(j) \leq \min \left\{ \begin{array}{l} n - k, \\ |\mathcal{L}''| + \overline{d_c}(j), \\ |\mathcal{L}''| + (|\mathcal{A}^k_{*j}| - \alpha_j) + \left(\sum_{e \in \mathcal{C}_j} |\mathcal{L}_e| \right). \end{array} \right.$$

This bound is similar to the bound used in the minimum degree ordering algorithm in MATLAB [21], except that it is used in a symmetric context and thus the diagonal entry is excluded from the summation. To compute this bound for all rows and columns in $\mathbf{C}$ would take time

$$\Theta \left(\sum_{i \in \mathcal{L}''} \beta_i + \sum_{j \in \mathcal{U}''} \alpha_j \right)$$

to scan $\mathcal{A}^k$ and time

$$\Theta \left(\sum_{i \in \mathcal{L}''} |\mathcal{R}_i| + \sum_{j \in \mathcal{U}''} |\mathcal{C}_j| \right)$$

to scan $\mathcal{R}_i$ and $\mathcal{C}_j$. For a single column j, the total time is $\Theta(\alpha_j + |\mathcal{C}_j|)$, or $O(|\mathcal{A}^k_{*j}| + |\mathcal{C}_j|)$, since $\alpha_j \leq |\mathcal{A}^k_{*j}|$. Similarly, the time to compute this loose degree bound for a row i is $\Theta(\beta_i + |\mathcal{R}_i|)$, or $O(|\mathcal{A}^k_{i*}| + |\mathcal{R}_i|)$.

However, a much tighter bound can be obtained in the *same* asymptotic time. The set $\mathcal{L}_e$ can be split into two disjoint subsets: the *external* subset $\mathcal{L}_e \setminus \mathcal{L}$ and the *internal* subset $\mathcal{L}_e \cap \mathcal{L}$, where $\mathcal{L}_e = (\mathcal{L}_e \setminus \mathcal{L}) \cup (\mathcal{L}_e \cap \mathcal{L})$, and "\" is the standard set difference operator. Define $|\mathcal{L}_e \setminus \mathcal{L}|$ as the *external column degree* of element e with respect to $\mathbf{F}$. Similarly, define $|\mathcal{U}_e \setminus \mathcal{U}|$ as the *external row degree* of element e with respect to $\mathbf{F}$. We use the bound

$$d_c(j) \le \overline{d_c}(j) = \min \begin{cases} n-k, \\ |\mathcal{L}''| + \overline{d_c}(j), \\ |\mathcal{L}''| + (|\mathcal{A}^k_{*j}| - \alpha_j) + \left(\sum_{e \in \mathcal{C}_j} |\mathcal{L}_e \setminus \mathcal{L}| \right), \end{cases} \tag{4.1}$$

which is tighter than before since $|\mathcal{L}_e \setminus \mathcal{L}| = |\mathcal{L}_e| - |\mathcal{L}_e \cap \mathcal{L}| \le |\mathcal{L}_e|$. The equation for $\overline{d_r}(i)$ is analogous.

An efficient way of computing the external row and column degrees is given in Algorithm 2. (The algorithm for external row degrees is analogous.) The array w is a work array of size n that is used to compute the external column degrees $|\mathcal{L}_e \setminus \mathcal{L}|$. We actually use a slight variation of Algorithm 2 that does not require the assumption that $w(e) = -1$.

ALGORITHM 2 (computation of external column degrees).

```
assume w(e) = −1, for all e ∈ V̄
for each new row i ∈ L do
    for each element e in the element list R_i of row i do
        if (w(e) < 0) then w(e) = |L_e|
        w(e) = w(e) − 1
    end for
end for
```

The cost of Algorithm 2 can be amortized over all subsequent degree updates on the current front. We use the term "amortized time" to define how much of this total work is ascribed to the computation of a single degree bound, $\overline{d_c}(j)$ or $\overline{d_r}(i)$. Note that in computing these amortized time estimates we actually include the cost of computing the external row degrees within the estimate for the column degree bounds although it is actually the external column degrees that are used in computing this bound. We can amortize the time in this way because we compute the external row and column degrees, and the row and column degree bounds, for all rows and columns in the current frontal matrix.

Relating our approximate degree algorithm to George and Liu's quotient graph, our algorithm takes an amortized time of $O(|\mathcal{A}^k_{*j}| + |\mathcal{C}_j|) = O(|\mathrm{Adj}_{\mathcal{G}_k}(x_j)|)$ to compute $\overline{d_c}(j)$. This correspondence holds only if $\mathcal{A}$ is symmetric and pivots are selected from the diagonal. This is much less than the $\Omega(|\mathrm{Adj}_{G_k}(x_j)|)$ time taken to compute the true degree. The true degree $d_c(j) = |\mathrm{Struct}(\mathbf{A}'_{*j})| = |\mathrm{Adj}_{G_k}(x_j)|$ is the degree of node x_j in the implicitly represented elimination graph, G_k [19]. If indistinguishable uneliminated nodes are present in the quotient graph (as used in [26], for example), both of these time complexity bounds are reduced, but computing the true degree still takes much more time than computing our approximate degree.

We now describe how we compute our degree bound $\overline{d_c}(j)$ in an amortized time of $O(|\mathcal{A}^k_{*j}| + |\mathcal{C}_j|)$. We compute the external column degrees by scanning each e in $\mathcal{R}_i$ for each "new" row i in $\mathcal{L}$, as shown in Algorithm 2. A row or column is new if it did not appear in $\mathcal{L}$ or $\mathcal{U}$ prior to the current pivot. Since $e \in \mathcal{R}_i$ implies $i \in \mathcal{L}_e$, row i must be internal (that is, $i \in \mathcal{L}_e \cap \mathcal{L}$). If Algorithm 2 scans element e, the term $w(e)$

is initialized to $|\mathcal{L}_e|$ and then decremented once for each internal row $i \in \mathcal{L}_e \cap \mathcal{L}$. In this case, at the end of Algorithm 2 three equivalent conditions hold:

1. e appears in the list $\mathcal{R}_i$ for some row i in $\mathcal{L}$,
2. the internal subset $\mathcal{L}_e \cap \mathcal{L}$ is not empty,
3. $w(e) = |\mathcal{L}_e| - |\mathcal{L}_e \cap \mathcal{L}| = |\mathcal{L}_e \setminus \mathcal{L}|$.

If Algorithm 2 did not scan element e in any $\mathcal{R}_i$, then the three following equivalent conditions hold:

1. e does not appear in the list $\mathcal{R}_i$ for any row i in $\mathcal{L}$,
2. the internal subset $\mathcal{L}_e \cap \mathcal{L}$ is empty,
3. $w(e) < 0$.

Combining these two cases, we obtain

$$|\mathcal{L}_e \setminus \mathcal{L}| = \left\{ \begin{array}{ll} w(e) & \text{if } w(e) \geq 0 \\ |\mathcal{L}_e| & \text{otherwise} \end{array} \right\} \text{ for all } e \in \overline{V}. \tag{4.2}$$

To compute the external row degrees of all elements, we scan the element list $\mathcal{C}_j$ for each new column j in $\mathcal{U}$ in an analogous manner (with a separate work array). The total time to compute both the external column degrees (Algorithm 2) and the external row degrees is $\Theta(\sum_{i \in \mathcal{L}''} |\mathcal{R}_i| + \sum_{j \in \mathcal{U}''} |\mathcal{C}_j|)$.

We now describe our combined degree update and numerical assembly phase. This phase uses the external row and column degrees for both the degree update and the numerical assembly. We compute $\overline{d_c}(j)$ and assemble elements by scanning the element list $\mathcal{C}_j$ for each column $j \in \mathcal{U}''$, evaluating $\overline{d_c}(j)$ using equations (4.1) and (4.2). If the external row and column degrees of element e are both zero, then we delete (e, m) from $\mathcal{C}_j$ and assemble $\mathbf{C}_e$ into $\mathbf{F}$. Element e no longer exists. This is identical to the assembly from a child (element e) into a parent (the current frontal matrix $\mathbf{F}$) in the assembly tree of the classical multifrontal method. It is also referred to as *element absorption* [13]. It is too costly at this point to delete all references to the deleted element. If a reference to a deleted element is found later on, it is then discarded. If the external column degree of element e is zero but its external row degree is not zero, then (e, m) is deleted from $\mathcal{C}_j$, column j is assembled from $\mathbf{C}_e$ into $\mathbf{F}$, and column j is deleted from element e. Finally, we scan the original entries $(\mathcal{A}^k_{*j})$ in column j as discussed in section 3.2. If all remaining entries can be assembled into the current frontal matrix, then we perform the assembly and delete column j of $\mathbf{A}^k$. Thus, the amortized time to compute $\overline{d_c}(j)$ is $O(|\mathcal{A}^k_{*j}| + |\mathcal{C}_j|)$. This time complexity does not include the time to perform the numerical assembly.

The scan of rows $i \in \mathcal{L}''$ is analogous. The amortized time to compute $\overline{d_r}(i)$ is $O(|\mathcal{A}^k_{i*}| + |\mathcal{R}_i|)$.

We use the sets $\mathcal{L}_e$ and $\mathcal{U}_e$ for all $e \in \overline{V}$ to represent the nonzero pattern of the active matrix using equations (3.5) and (3.7). Our combined degree update and numerical assembly phase reduces the storage required for this representation. These reductions are summarized below:

1. If $|\mathcal{L}_e \setminus \mathcal{L}| = 0$ and $|\mathcal{U}_e \setminus \mathcal{U}| = 0$ then all of $\mathbf{C}_e$ is assembled into $\mathbf{F}$. Element e and all entries in $\mathcal{L}_e$ and $\mathcal{U}_e$ are deleted. This is the same as the complete element absorption that occurs in the classical multifrontal method. Symbolically, this is also the same as element absorption in a quotient graph-based minimum degree ordering algorithm.
2. If $|\mathcal{L}_e \setminus \mathcal{L}| = 0$ and $|\mathcal{U}_e \setminus \mathcal{U}| \neq 0$ then columns $\mathcal{U}_e \cap \mathcal{U}$ are assembled from $\mathbf{C}_e$ into $\mathbf{F}$. The entries $\mathcal{U}_e \cap \mathcal{U}$ are deleted from $\mathcal{U}_e$.
3. If $|\mathcal{L}_e \setminus \mathcal{L}| \neq 0$ and $|\mathcal{U}_e \setminus \mathcal{U}| = 0$ then rows $\mathcal{L}_e \cap \mathcal{L}$ are assembled from $\mathbf{C}_e$ into

F. The entries $\mathcal{L}_e \cap \mathcal{L}$ are deleted from $\mathcal{L}_e$. An example of this assembly is discussed below.

For pivot a'_{22} in the example, we only have one previous element, element 1. The element lists are shown in Table 3.2. The external column degree of element 1 is one, since $|\mathcal{L}_1| = 4$, and $e = 1$ appears in the element lists of three rows in $\mathcal{L}$. The external row degree of element 1 is zero, since $|\mathcal{U}_1| = 2$, and $e = 1$ appears in the element lists of two columns in $\mathcal{U}$. We have $\mathcal{L}_1 = (\mathcal{L}_1 \setminus \mathcal{L}) \cup (\mathcal{L}_1 \cap \mathcal{L}) = \{4\} \cup \{2, 3, 7\}$ and $\mathcal{U}_1 = (\mathcal{U}_1 \setminus \mathcal{U}) \cup (\mathcal{U}_1 \cap \mathcal{U}) = \emptyset \cup \{4, 5\}$. Rows 2, 3, and 7 (but not 4) are assembled from $\mathbf{C}_1$ into $\mathbf{F}$ and deleted. This reduction and assembly corresponds to case 3, above. Row 2 and columns 2 and 3 of $\mathbf{A}^k$ are also assembled into $\mathbf{F}$. No columns are assembled from $\mathbf{C}_1$ into $\mathbf{F}$ during the column scan, since the external column degree of element 1 is not zero.

We have

$$\mathbf{C}_1 = \left\{ \begin{array}{c|cc} & 4 & 5 \\ \hline & - & - \\ 4 & - & - \\ & c_{44} & c_{45} \\ & - & - \end{array} \right\}, \quad \mathbf{A}^k = \left\{ \begin{array}{c|ccc} & 5 & 6 & 7 \\ \hline 3 & 0 & 0 & a_{37} \\ 4 & a_{45} & a_{46} & 0 \\ 5 & a_{55} & a_{56} & 0 \\ 6 & 0 & a_{66} & a_{67} \\ 7 & a_{75} & 0 & a_{77} \end{array} \right\},$$

and

$$\mathbf{F} = \left\{ \begin{array}{c|c|c} & \mathcal{U}' & \mathcal{U}'' \\ \hline \mathcal{L}' & a'_{rc} & \mathbf{A}'_{r*} \\ \hline \mathcal{L}'' & \mathbf{A}'_{*c} & \mathbf{C} \end{array} \right\} = \left\{ \begin{array}{c|c|ccc} & 2 & 3 & 4 & 5 \\ \hline 2 & a'_{22} & a'_{23} & a'_{24} & a'_{25} \\ \hline 3 & a'_{32} & a_{33} & c_{34} & c_{35} \\ 5 & a'_{52} & a_{53} & 0 & 0 \\ 7 & a'_{72} & 0 & c_{74} & c_{75} \end{array} \right\},$$

where we have marked already assembled parts of element 1 by $-$. The set $\mathcal{L}_1$ is now only $\{4\}$, the other entries (2, 3, and 7) having been deleted. It would be possible to recover this space during the computation but we have chosen not to do so in the interest of avoiding the expense of updating the associated element lists. Note then that these lists refer to positions within the original element.

The assembly and deletion of a row in an element does not affect the external column degree of the element, which is why only new rows are scanned in Algorithm 2. Similarly, the assembly and deletion of a column in an element does not affect the external row degree of the element.

The local pivot search within $\mathbf{F}$ evaluates the candidate column c and row r using equations (3.4), (3.5), and (3.7). In the example, the second pivot a'_{33} is found in the local pivot search. The set $\mathcal{L}$ remains unchanged, but the set $\mathcal{U}$ is augmented with the new column 7. Rows 3 and 7 are assembled from $\mathbf{A}^k$ into $\mathbf{F}$ in the subsequent execution of step 2 for this pivot. No further assembly from $\mathbf{C}_1$ is made.

Step 2 is substantially reduced if there are no new rows or columns in $\mathbf{F}$. No assemblies from $\mathbf{A}^k$ or $\mathbf{C}_e$ can be done since all possible assemblies would have been done for a previous pivot. It is only necessary to decrement $\overline{d_c}(j)$ for all $j \in \mathcal{L}''$ and $\overline{d_r}(i)$ for all $i \in \mathcal{U}''$.

4.3. Step 3: Complete the factorization of F. In the example, the final factorized frontal matrix is

$$\mathbf{F} = \left\{ \begin{array}{c|c} & \begin{array}{cc} \mathcal{U}' & \mathcal{U}'' \end{array} \\ \hline \begin{array}{c} \mathcal{L}' \\ \mathcal{L}'' \end{array} & \left[\begin{array}{c|c} \mathbf{L}'\mathbf{U}' & \mathbf{U}'' \\ \hline \mathbf{L}'' & \mathbf{C} \end{array} \right] \end{array} \right\} = \left\{ \begin{array}{c|ccccc} & 2 & 3 & 4 & 5 & 7 \\ \hline 2 & u_{22} & u_{23} & u_{24} & u_{25} & 0 \\ 3 & l_{32} & u_{33} & u_{34} & u_{35} & u_{37} \\ \hline 5 & l_{52} & l_{53} & c_{54} & c_{55} & c_{57} \\ 7 & l_{72} & l_{73} & c_{74} & c_{75} & c_{77} \end{array} \right\}.$$

Note that $u_{27} = 0$, due to the relaxed amalgamation of two pivot rows with non-identical patterns. Relaxed amalgamation can result in higher performance since more of the Level 3 BLAS can be used. In the small example, the active matrix is represented by the implicit assembly

$$\begin{aligned} \mathbf{A}' &= \mathbf{C}_1 \dotplus \mathbf{C}_2 \dotplus \mathbf{A}^k \\ &= \left\{ \begin{array}{c|cc} & 4 & 5 \\ \hline 4 & c_{44} & c_{45} \end{array} \right\} \dotplus \left\{ \begin{array}{c|ccc} & 4 & 5 & 7 \\ \hline 5 & c_{54} & c_{55} & c_{57} \\ 7 & c_{74} & c_{75} & c_{77} \end{array} \right\} \\ &\quad \dotplus \left\{ \begin{array}{c|ccc} & 5 & 6 & 7 \\ \hline 4 & a_{45} & a_{46} & 0 \\ 5 & a_{55} & a_{56} & 0 \\ 6 & 0 & a_{66} & a_{67} \end{array} \right\} \\ &= \left\{ \begin{array}{c|cccc} & 4 & 5 & 6 & 7 \\ \hline 4 & a'_{44} & a'_{45} & a'_{46} & 0 \\ 5 & a'_{54} & a'_{55} & a'_{56} & a'_{57} \\ 6 & 0 & 0 & a'_{66} & a'_{67} \\ 7 & a'_{74} & a'_{75} & 0 & a'_{77} \end{array} \right\}. \end{aligned}$$

The element lists are shown in Table 4.1.

TABLE 4.1
Element lists for example matrix after second frontal matrix.

i	$\mathcal{R}_i$	j	$\mathcal{C}_j$
4	(1,3)	4	(1,1) (2,1)
5	(2,1)	5	(1,2) (2,2)
6	-	6	-
7	(2,2)	7	(2,3)

5. Algorithm. Algorithm 3 is a full outline of the UMFPACK (Version 1.0) algorithm.

ALGORITHM 3 (unsymmetric-pattern multifrontal algorithm).

0: initializations
$k = 1$
$\overline{V}$ = empty
while $(k \leq n)$ **do**
1: $e = k$

```
        global search for kth pivot: a'_rc
        L = Struct(A'_*c)
        U = Struct(A'_r*)
        s = g|L|
        t = g|U|
        form rectangular frontal matrix F in an s-by-t work array
        do until an exit condition (marked with **) is satisfied
2:          assembly and degree update:
            assemble kth pivot row and column into F
            scan element lists and compute external degrees
            assemble rows and columns from A^k into F
            assemble contribution blocks into F
            compute degree bounds
            numerical update:
            compute entries of L (F_*c = F_*c / a'_rc)
            k = k + 1
            if (|U'| mod b = 0) C = C - L̂Û
            local pivot search and numerical update of candidates:
**          if (|U''| = 0) exit this loop
            find candidate pivot column c ∈ U''
            C_*c = C_*c - L̂Û_*c
            if (d̄_c(c) ≠ |L''|) assemble column c and compute d_c(c)
**          if (d_c(c) > s - |U'|) exit this loop
            find candidate pivot row r ∈ L''
**          if (r not found) exit this loop
            if (d̄_r(r) ≠ |U''|) assemble row r and compute d_r(r)
**          if (d_r(r) > t - |U'|) exit this loop
            L = L ∪ Struct(A'_*c)
            U = U ∪ Struct(A'_r*)
            C_r* = C_r* - L̂_r*Û
        enddo
3:      final numerical update and saving of contribution block, C:
        save L', L'', L, U', U'', and U
        C = C - L̂Û
        C_e = C
        place C_e in heap
        L_e = L''
        U_e = U''
        delete F
        V̄ = V̄ ∪ {e}
        add e to element lists
    endwhile
```

6. Performance results. In this section, we compare the performance of UMFPACK Version 1.0 with MUPS [2] and MA48 [15] on a single processor of a CRAY C-98 (although MUPS is a parallel code). Each method has a set of input parameters that controls its behavior. We used the recommended defaults for most of these, with a few exceptions that we indicate below. All methods can factorize general unsymmetric matrices, and all use dense matrix kernels to some extent [6].

TABLE 6.1
Test matrices.

name	n	$\lvert\mathbf{A}\rvert$	sym.	discipline	comments
GRE 1107	1107	5664	0.000	discrete simul.	computer system
GEMAT11	4929	33185	0.001	electric power	linear programming basis
ORANI678	2529	90158	0.071	economics	Australia
PSMIGR 1	3140	543162	0.479	demography	US county-to-county migration
LNS 3937	3937	25407	0.850	fluid flow	linearized Navier–Stokes
HYDR1	5308	23752	0.004	chemical eng.	dynamic simulation
RDIST1	4134	94408	0.059	chemical eng.	reactive distillation
LHR04	4101	82682	0.015	chemical eng.	light hydrocarbon recovery
LHR71	70304	1528092	0.002	chemical eng.	light hydrocarbon recovery

MA48 [15] supersedes the MA28 code [12]. It first performs an ordering phase that also computes most of the factors but discards them. It then performs the numerical factorization to compute the entire LU factors. When the matrix becomes dense enough near the end of factorization (default of 50% dense), MA48 switches to a dense factorization code.

MUPS performs a minimum degree ordering and symbolic factorization on the nonzero pattern of $\mathbf{A} + \mathbf{A}^T$ and constructs an assembly tree for the numerical factorization phase [2, 8, 9, 14]. During numerical factorization, candidate pivot entries must pass a threshold partial pivoting test similar to equation (3.1), except that the test is by rows instead of by columns. Since the other methods we are comparing perform this test by columns, we factorize $\mathbf{A}^T$ with MUPS and then use the factors of $\mathbf{A}^T$ to solve the original system ($\mathbf{Ax} = \mathbf{b}$). MUPS optionally preorders a matrix so that the diagonal is zero free using a maximum transversal algorithm [7]. MUPS always attempts to preserve symmetry. It does not permute the matrix to block upper triangular form. Note that we do not include symmetric-patterned matrices in our test set, for which MUPS is nearly always faster than UMFPACK.

By default, both UMFPACK and MA48 preorder a matrix to block upper triangular form (always preceded by finding a maximum transversal [7]) and then factorize each block on the diagonal [11]. Off-diagonal blocks do not suffer fill in. This can reduce the work for unsymmetric matrices. We did not perform this preordering, since MUPS does not provide the option. UMFPACK has similar input parameters to MA48, although it does not explicitly include a switch to dense factorization code (each frontal matrix is dense, however). We selected the threshold partial pivoting factor (u) to be 0.1 for all four methods.

The methods were tested on a single processor of a CRAY C-98, with 512 Megawords of memory (8-byte words). Version 6.0.4.1 of the Fortran compiler (CFT77) was used. Each method was given 95Mw of memory to factorize the test matrices, listed in Table 6.1. The table lists the name, order, number of entries ($|\mathbf{A}|$), symmetry, the discipline from which the matrix came, and additional comments. The symmetry is the number of *matched* off-diagonal entries over the total number of off-diagonal entries. An entry, a_{ij} ($j \neq i$), is matched if a_{ji} is also an entry. All matrices are available via anonymous ftp. They include matrices from the Harwell–Boeing Collection [10]. One matrix (LHR71) was so ill conditioned that it required scaling prior to its factorization. The scale factors were computed by the Harwell Subroutine Library routine MC19A [3]. Each row was then subsequently divided by the maximum absolute value in the row (or column, depending on how the method implements threshold partial pivoting). No scaling was performed on the other matrices.

The results are shown in Table 6.2. For each matrix, Table 6.2 lists the numerical

TABLE 6.2
Results.

Matrix	method	factor (sec)	total (sec)	$\|\mathbf{L}+\mathbf{U}\|$ (10^6)	memory (10^6)	op count (10^6)
GRE 1107	UMF.	.07	0.30	.09	.3	9.7
	MA48	.11	0.38	.07	.3	8.1
	MUPS	.13	0.38	.19	.4	26.6
GEMAT11	UMF.	.18	.45	.08	.4	1.0
	MA48	.18	.54	.05	.4	.7
	MUPS	.27	.57	.14	.4	2.8
ORANI678	UMF.	.53	2.07	.12	1.1	7.4
	MA48	.32	1.01	.15	.8	14.2
	MUPS	.61	218.69	.39	13.3	87.6
PSMIGR 1	UMF.	15.62	33.99	6.36	26.4	10194.8
	MA48	14.92	28.86	6.40	20.9	10465.3
	MUPS	14.04	323.15	6.21	26.9	9002.4
LNS 3937	UMF.	.45	1.89	.50	1.4	84.8
	MA48	1.00	3.37	.69	2.2	280.4
	MUPS	.71	1.73	.92	1.2	185.8
HYDR1	UMF.	.24	1.05	.15	.6	4.5
	MA48	.28	.81	.08	.4	.9
	MUPS	.57	1.21	.24	.5	10.7
RDIST1	UMF.	.47	1.53	.49	1.4	37.1
	MA48	1.37	4.78	.41	1.6	27.2
	MUPS	.33	2.01	.28	.7	10.3
LHR04	UMF.	.56	2.51	.39	1.5	30.6
	MA48	1.27	4.25	.34	1.3	25.8
	MUPS	1.03	9.89	1.10	2.3	300.3
LHR71	UMF.	12.26	53.80	10.49	30.2	1294.5
	MA48	51.60	171.66	10.08	36.1	1338.4
	MUPS	-	-	-	> 95.0	-

factorization time, total factorization time, number of nonzeros in $\mathbf{L}+\mathbf{U}$ (in millions), amount of memory used (in millions of words), and floating-point operation count (in millions of operations) for each method. The total time includes preordering, symbolic analysis and factorization, and numerical factorization. The time to compute the scale factors for the LHR71 matrix is not included, since we used the same scaling algorithm for all methods. For each matrix, the lowest time, memory usage, or operation count is underlined. We compared the solution vectors, $\mathbf{x}$, for each method. We found that all four methods compute the solutions with comparable accuracy, in terms of the norm of the residual. We do not give the residual in Table 6.2.

MUPS failed on the LHR71 matrix because of insufficient memory. This is a very ill-conditioned problem that causes MUPS to be unable, on numerical grounds, to choose pivots selected by the analysis. This leads to an increase in fill in and subsequent failure.

We also compared UMFPACK with Gilbert and Peierls' partial pivoting code, GPLU [22], and with SSGETRF, a classical multifrontal method in the CRAY Research Library. The peak performance of GPLU was low primarily because its innermost loops do not readily vectorize (even with the appropriate compiler directives). Since this is a limitation of the code and not a fundamental limitation of the algorithm, we do not report the GPLU results. The results for SSGETRF were roughly comparable with MUPS, except that SSGETRF tended to use slightly less memory than MUPS (sometimes as little as 65% of that of MUPS) and was typically slightly slower than MUPS (sometimes twice as slow as MUPS). We thus do not report

the SSGETRF results since they do not change our overall comparison between the classical multifrontal method (MUPS or SSGETRF) and our unsymmetric-pattern multifrontal method (UMFPACK).

Overall, these results show that the unsymmetric-pattern multifrontal method is a competitive algorithm when compared with the classical multifrontal approach (MUPS) and an algorithm based on more conventional sparse matrix data structures (MA48).

Acknowledgments. We thank Patrick Amestoy, Mario Arioli, Michel Daydé, Theodore Johnson, and Steve Zitney for many helpful discussions; John Gilbert for providing a copy of the GPLU factorization code; and Joseph Liu for providing a copy of the MMD ordering code (used for GPLU). Many researchers provided us with large unsymmetric matrices, a class of matrices that is weak in Release 1 of the Harwell–Boeing Collection. We would also like to thank the referees, whose comments and suggestions improved the presentation of the paper.

REFERENCES

[1] P. 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] P. R. Amestoy and I. S. Duff, *Vectorization of a multiprocessor multifrontal code*, Internat. J. Supercomputer Appl., 3 (1989), pp. 41–59.

[3] A. R. Curtis and J. K. Reid, *On the automatic scaling of matrices for Gaussian elimination*, J. Inst. Math. Appl., 10 (1972), pp. 118–124.

[4] T. A. Davis, *Users' Guide to the Unsymmetric-Pattern Multifrontal Package (UMFPACK, Version* 1.0*)*, Technical Report TR-93-020, CISE Dept., Univ. of Florida, Gainesville, FL, 1993. Version 1.0 has been superseded by Version 2.0. For a copy of Version 2.0, send e-mail to netlib@ornl.gov with the one-line message send index from linalg.

[5] T. A. Davis and I. S. Duff, *Unsymmetric-Pattern Multifrontal Methods for Parallel Sparse LU Factorization*, Technical Report TR-91-023, CISE Dept., Univ. of Florida, Gainesville, FL, 1991.

[6] J. J. Dongarra, J. J. Du Croz, I. S. Duff, and S. Hammarling, *A set of level* 3 *basic linear algebra subprograms*, ACM Trans. Math. Software, 16 (1990), pp. 1–17.

[7] I. S. Duff, *On algorithms for obtaining a maximum transversal*, ACM Trans. Math. Software, 7 (1981), pp. 315–330.

[8] I. S. Duff, *Parallel implementation of multifrontal schemes*, Parallel Comput., 3 (1986), pp. 193–204.

[9] I. S. Duff, A. M. Erisman, and J. K. Reid, *Direct Methods for Sparse Matrices*, Oxford Univ. Press, London, 1986.

[10] I. S. Duff, R. G. Grimes, and J. G. Lewis, *Users' Guide for the Harwell-Boeing Sparse Matrix Collection (Release* 1*)*, Technical Report RAL-92-086, Rutherford Appleton Laboratory, Didcot, Oxon, England, 1992.

[11] I. S. Duff and J. K. Reid, *An implementation of Tarjan's algorithm for the block triangularization of a matrix*, ACM Trans. Math. Software, 4 (1978), pp. 137–147.

[12] I. S. Duff and J. K. Reid, *Some design features of a sparse matrix code*, ACM Trans. Math. Software, 5 (1979), pp. 18–35.

[13] I. S. Duff and J. K. Reid, *The multifrontal solution of indefinite sparse symmetric linear systems*, ACM Trans. Math. Software, 9 (1983), pp. 302–325.

[14] I. S. Duff and J. K. Reid, *The multifrontal solution of unsymmetric sets of linear equations*, SIAM J. Sci. Comput., 5 (1984), pp. 633–641.

[15] I. S. Duff and J. K. Reid, *MA*48*, a Fortran Code for Direct Solution of Sparse Unsymmetric Linear Systems of Equations*, Technical Report RAL-93-072, Rutherford Appleton Laboratory, Didcot, Oxon, England, 1993.

[16] S. C. Eisenstat and J. W. H. Liu, *Exploiting structural symmetry in unsymmetric sparse symbolic factorization*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 202–211.

[17] S. C. Eisenstat and J. W. H. Liu, *Exploiting structural symmetry in a sparse partial pivoting code*, SIAM J. Sci. Statist. Comput., 14 (1993), pp. 253–257.

[18] S. C. Eisenstat and J. W. H. Liu, *Structural representations of Schur complements in sparse matrices*, in Graph Theory and Sparse Matrix Computation, The IMA Volumes in Mathematics and its Applications, Vol. 56, A. George, J. R. Gilbert, and J. W. H. Liu, eds., Springer-Verlag, Berlin, 1993, pp. 85–100.

[19] A. George and J. W. H. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[20] J. R. Gilbert and J. W. H. Liu, *Elimination structures for unsymmetric sparse LU factors*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 334–354.

[21] 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.

[22] J. R. Gilbert and T. Peierls, *Sparse partial pivoting in time proportional to arithmetic operations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 862–874.

[23] S. M. Hadfield, *On the LU Factorization of Sequences of Identically Structured Sparse Matrices within a Distributed Memory Environment*, Ph.D. thesis, CISE Dept., Univ. of Florida, Gainesville, FL, 1994; Technical Report TR-94-019, CISE Dept., Univ. of Florida, Gainesville, FL.

[24] S. M. Hadfield and T. A. Davis, *Potential and achievable parallelism in the unsymmetric-pattern multifrontal LU factorization method for sparse matrices*, in Proc. Fifth SIAM Conf. on Applied Linear Algebra, Snowbird, UT, 1994, SIAM, Philadelphia, PA.

[25] S. M. Hadfield and T. A. Davis, *Lost pivot recovery for an unsymmetric-pattern multifrontal method*, Technical Report TR-94-029, CISE Dept., Univ. of Florida, Gainesville, FL.

[26] J. W. H. Liu, *Modification of the minimum-degree algorithm by multiple elimination*, ACM Trans. Math. Software, 11 (1985), pp. 141–153.

[27] J. W. H. Liu, *The multifrontal method for sparse matrix solution: Theory and practice*, SIAM Review, 34 (1992), pp. 82–109.

[28] H. M. Markowitz, *The elimination form of the inverse and its application to linear programming*, Management Sci., 3 (1957), pp. 255–269.

[29] Z. Zlatev, *On some pivotal strategies in Gaussian elimination by sparse technique*, SIAM J. Numer. Anal., 17 (1980), pp. 18–30.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 159–180, January 1997

013

SPARSE MULTIFRONTAL RANK REVEALING *QR* FACTORIZATION*

DANIEL J. PIERCE† AND JOHN G. LEWIS†

Abstract. We describe an algorithm to compute an approximate rank revealing sparse QR factorization. We use a two phase algorithm to provide especially high accuracy in the labeling of some columns as "redundant," which ensures robustness in the use of our factorization in computing explicit bases of the nullspace.

Our first phase is similar in outline to other proposed sparse $RRQR$ factorizations, in that we couple a standard sparse QR factorization scheme with a condition estimator to develop a factorization with a well-conditioned leading block. There are important details in our implementation of the condition estimator and pivoting that enhance efficiency and reliability. However, the exceptional characteristic of our algorithm is its second phase, which ensures that columns labeled as redundant lead to highly accurate nullvectors. The second phase requires that we compute all columns of R explicitly in the first phase; we cannot discard "redundant" columns as is often done elsewhere. This condition, in the presence of pivoting to reveal the rank, requires dynamic data structures and necessarily degrades sparsity. But the additional work fits naturally into the multifrontal factorization's use of efficient dense vector kernels, minimizing overall cost.

We present a theoretical analysis that shows that our use of approximate singular vectors does not degrade the quality of our rank-revealing factorization; we achieve an exponential bound like methods that use exact singular vectors. We provide results of numerical experiments and close with a discussion of limitations of this approach.

Key words. multifrontal, rank revealing, sparse matrix, least squares, rank deficient

AMS subject classifications. 15A03, 15A06, 15A23, 65F20, 65F50

PII. S0895479893244353

1. Introduction.

This paper describes an algorithm to compute a rank revealing QR (an $RRQR$) factorization of a sparse overdetermined matrix. Such a factorization is useful for removing redundant constraints in optimization or for handling very ill-conditioned linear systems. Rank-deficient geometric design problems are common in aerospace applications [14]; the $RRQR$ factorization provides a means to perform additional optimization of the aircraft part beyond merely satisfying geometric criteria. Chan and Hansen [10] survey a wide variety of applications. Some of our applications require explicit knowledge of the derived nullspace basis, which can easily be computed from our factorization.

An $RRQR$ factorization of the $m \times n$ matrix A, where $m \geq n$, is an orthogonal factorization of the form

$$Q^T A P = \begin{bmatrix} R & S \\ 0 & T \end{bmatrix}, \tag{1}$$

where Q is orthogonal, P is a permutation matrix, the condition number of R is not large, and $\|T\|_2$ is small. Hong and Pan [24] show that such a factorization always exists, but their scheme for computing the permutation is not computable in a reasonable number of operations.

Practical $RRQR$ algorithms use heuristics to choose permutations that give approximate solutions that usually do not guarantee that both of the $RRQR$ conditions

* Received by the editors February 16, 1993; accepted for publication (in revised form) by J. W. H. Liu February 3, 1996.
http://www.siam.org/journals/simax/18-1/24435.html
† Boeing Information and Support Services, P.O. Box 3707, MS 7L-22, Seattle, WA 98124-2207 (dpierce@atc.boeing.com, jglewis@espresso.rt.cs.boeing.com).

hold equally. Chandrasekaran and Ipsen [9] categorize such *RRQR* algorithms by whether they make R well conditioned (Type I) or they make the norm of T small (Type II). More precisely these conditions on R and T define two separate problems:

$$\text{Problem I:} \quad \max_P \sigma_{\min}(R),$$
$$\text{Problem II:} \quad \min_P \sigma_{\max}(T).$$

That is, Problem I corresponds to selecting the most linearly independent columns of A to compute R, while Problem II corresponds to selecting the most linearly dependent columns in A as the second block column. It is not known [9] whether there must exist a common solution to both problems, which is a stronger condition than an *RRQR* factorization.

In practice, it is necessary to prioritize between the conditioning of R and the size of $\|T\|_2$, and to balance both against the factorization cost. Pan and Tang [32] present a different set of techniques which produce *RRQR* factorizations that try to satisfy

$$\sigma_{\min}(R) \approx \sigma_k(A) \qquad \text{and} \qquad \sigma_{\max}(T) \approx \sigma_{k+1}(A).$$

The techniques in both [9] and [32] require iterative interchanging of columns between R and the column set that defines the T matrix. (The latter set are the so-called "redundant" columns.)

Maintaining sparsity in the factorization adds another dimension of complexity. Any extensive use of interchanges between R and the redundant columns usually causes a rapid degradation of sparsity. Consequently previous sparse algorithms [1, 19, 23] have effectively permanently discarded any columns that are ever labeled as redundant. Such algorithms can be implemented with static data structures because the storage for the columns chosen in R is known to be a subset of the storage needed for a standard unpivoted QR factorization.

In our applications it is particularly important that the derived nullspace bases be accurate nullvectors. Thus, it is essential that we approximate solutions to Problem II well. Yet, sparsity constrains us to very limited interchanges between R and the redundant columns. We choose an algorithm with two phases. Phase I computes an initial factorization of the form (1) so that the condition of R is less than a user specified tolerance, $1/\tau$. Phase II reinstates columns of the matrix

$$\begin{bmatrix} S \\ T \end{bmatrix}$$

into the factor R to ensure that $\|T\|_2$ is small enough. The first pass of our algorithm is a Type I method. However, our method allows columns that were placed into the R factor to be removed later, similar to Foster's Algorithm 2 [15]. The second phase of the algorithm, which ensures that $\|T\|_2$ is small, is a Type II algorithm. We do not iterate this process because of the potential dramatic effects on the sparsity of the matrix and increases in overall computational requirements. In practice we determine a reasonably conditioned R and a T with small norm, without iterating.

Our algorithm differs from other sparse *RRQR* algorithms, most fundamentally in using dynamic data structures. This permits our second phase, which is not possible in the context of static data structures. In having a second phase that reinstates columns, our algorithm is unique among sparse codes. Our first phase is similar to the Type I algorithms in [1, 19, 23] but differs in the choice of heuristic to estimate the condition of the partial factorization R. We use the sparse variant of incremental condition estimation [4], which provides an estimate in the Euclidean norm; the

other algorithms use different norms with less direct connection to the singular value property needed here. In addition, our condition estimate considers all of the partial factorization at each step. We are free at any step to choose any column in the partial factorization as redundant. Condition estimators are usually vulnerable to very special counterexamples, but in practice each of these estimators works well in practice. Our scheme has additional advantages in reducing cost by permitting all Type I pivots to occur as soon as possible.

Our algorithm uses the multifrontal QR algorithm as the basis of the factorization. We assume the reader has some familiarity with this sparse QR algorithm, first introduced by Liu [28]; other references include [18], [26], [31]. The multifrontal algorithm was chosen for efficiency. It has a low operation count, compared with the classical work of George and Heath [16], and it organizes the computations to take advantage of dense matrix computational kernels [11], [26], [28]. We are able to exploit the multifrontal paradigm even as we allow dynamic changes in the factor R.

The columns of A are initially reordered to preserve sparsity, using the structure of A^TA, as first proposed in [16]. The QR factorization is computed one row at a time. At the same time we estimate the condition of the triangular portion of the matrix computed thus far. A rank deficiency is signaled by the condition estimate exceeding a tolerance, in which case we remove the column in the computed factor that we estimate is the most linearly dependent and update the triangular factor.

We estimate the condition number of a sparse submatrix with SPICE [4] (SParse Incremental Condition Estimation), a generalization of the incremental condition estimation scheme ICE [2]. SPICE fits quite naturally into the multifrontal scheme, which allows us to save storage and localize computations. Localization is important for sparse problems because accessing global information in a sparse matrix can be very expensive, especially if the matrix is stored out of core.

We estimate the most linearly dependent column by performing a back solve with the partially computed triangular factor. This operation can often be confined to a subset of the computed columns of the factor through the sparsity of R. We use dense Givens rotations to update the factorization when a column is removed from R, where again sparsity can be used to reduce the work and data access. The rest of the factorization is accomplished with dense Householder transformations, as in the standard multifrontal QR factorization.

If the right-hand side is known, we can apply the orthogonal transformations to it during the factorization. A back solve will complete the least squares solution. We do not save the orthogonal transformations, so we solve the seminormal equations when the solutions for additional right-hand sides are required. The details of corrected seminormal equations for rank deficient least squares problems are given by Pierce [35].

In section 2 we provide a basic explanation of the multifrontal algorithm for computing the QR factorization of a sparse matrix. Section 3 describes SPICE, the cornerstone to Phase I of the sparse $RRQR$ method. Section 4 describes Phase I and the pivoting scheme for removing columns from the computed factor. This produces a well-conditioned factor R. In section 5 we prove that the norm of the submatrix T is bounded above by a number that depends exponentially on the order of T. When the order of T is large, we need a second phase for the algorithm, to guarantee that $\|T\|_2$ is small. The second phase is discussed in section 6. Section 7 describes some preliminary test results with the algorithm. In the final section we describe additional techniques that could make the method even more robust.

2. The sparse multifrontal *QR* factorization. The sparse multifrontal QR factorization has two major advantages over more traditional sparse QR factorizations. Its operation count is usually lower than other schemes, and these fewer operations are grouped in ways that allow use of faster dense matrix/vector operations such as in [6]. The reduced operation count stems from *row merging* or *submatrix rotations*, notions introduced by Liu [28]. Reorganizing the data and computations to use dense vector operations uses the same paradigms as the algorithms developed by Duff and Reid for computing sparse Cholesky, symmetric indefinite LDL^T, and LU [12, 13] factorizations. Liu introduced the multifrontal QR factorization in [28], described there as the row merge tree QR factorization.

The row merge idea, which permeates the sparse factorization, is simple to explain through a dense example. Suppose that we have computed the QR factorization of a large, dense, overdetermined matrix. We would like to update this factorization to account for 20 new, sparse, data rows, each of which has nonzeros only in a common set of 5 columns. We could update the factorization by treating each new data row separately, using Givens rotations and requiring $\mathcal{O}(n^2)$ operations for each row. Overall this would require $\mathcal{O}(20n^2)$ operations.

The row merge alternative begins by computing the QR factorization of the 20 new rows. The resulting 5×5 triangular factor, the orthogonal reduction of the original 20 rows, can then be merged row by row into the large triangular factor. Givens rotations for this merger would require only $\mathcal{O}(5n^2)$ operations in total. The key is that the orthogonal reduction of a set of rows is equivalent to the original rows in terms of the overall orthogonal factorization.

Row merging is not limited to updating the factorization; it can be used throughout the factorization process. Generically each step in a row merge factorization takes as input a set of rows, some of which may be original data rows of A and others may be the results of earlier merge steps. The output from a row merge step is a small orthogonal factorization that gives the orthogonal reduction of the input rows.

Row merging introduces a new problem: choosing how to group the data rows to obtain the optimal reduction in work. This is apparently a very difficult problem; current algorithms use heuristics that are generally effective but not necessarily optimal.

A very simple heuristic is a columnwise algorithm, in which the factorization is completed in n row merge steps. At the ith step, collect all rows of the transformed matrix $\hat{A}$ (initially A) that have zeros in their first $i-1$ columns and a nonzero in their ith column. Compute the orthogonal reduction of these rows ($\hat{A} \leftarrow Q^T\hat{A}$). Then the first row of the orthogonal reduction of this set of rows is the ith row of the final factor R. The other rows of the orthogonal reduction must be processed further at later steps. That the first row is the ith row of R follows from the fact that all remaining unprocessed rows are entirely zero in their first i columns. For consistency with the original multifrontal algorithms, we describe the other rows of the orthogonal reduction as "generated data rows." Each row merge step for this algorithm has as input original and generated data rows (i.e., rows of $\hat{A}$ that have not been transformed by any orthogonal reductions, and those that have been); its output is the ith row of R and a new set of generated data rows.

The multifrontal QR factorization is a more sophisticated version of this simple heuristic, which uses a tree structure in the sparse factor R to better organize the data and operations and to use row merging more completely. In the remainder of this section we provide a brief outline of the multifrontal QR factorization. For more details the reader is referred to [26]. We assume the reader is familiar with basic graph

theoretic terms associated with sparse matrix computations, such as graphs of matrices, adjacency structures, and trees. (George and Liu [17] give a good introduction to these ideas.)

2.1. The basic multifrontal QR algorithm. The key to the organization of the sparse multifrontal QR algorithm is the elimination tree of the triangular factor R, which is defined by the sparsity pattern of R.

DEFINITION 2.1 (elimination tree of R). *Given an $m \times n$ matrix A ($m \geq n$) such that A^TA is irreducible, let R denote the upper triangular factor of the QR factorization of A. The elimination tree of A^TA and of R is a tree of n nodes, labeled $\{1, 2, 3, \ldots, n\}$, where the tree structure is defined by the condition that j is the parent of node i in the elimination tree if and only if $i < j$ and r_{ij} is the first off-diagonal nonzero in the ith row of R. (This condition defines an elimination forest when A^TA is irreducible.)*

For further discussion and insights regarding the elimination tree see [30]. Consider the matrix A in Figure 1, whose factor R is given in Figure 2 and whose corresponding elimination tree is given in Figure 3. The sparsity structure of the matrix A arises from a nested dissection ordering of the 3×3 model problem in [28].

$$\begin{bmatrix}
a_{1,1} & & a_{1,3} & & & & a_{1,7} & & a_{1,9} \\
a_{2,1} & & a_{2,3} & & & & a_{2,7} & & a_{2,9} \\
a_{3,1} & & a_{3,3} & & & & a_{3,7} & & a_{3,9} \\
a_{4,1} & & a_{4,3} & & & & a_{4,7} & & a_{4,9} \\
 & a_{5,2} & a_{5,3} & & & & & a_{5,8} & a_{5,9} \\
 & a_{6,2} & a_{6,3} & & & & & a_{6,8} & a_{6,9} \\
 & a_{7,2} & a_{7,3} & & & & & a_{7,8} & a_{7,9} \\
 & a_{8,2} & a_{8,3} & & & & & a_{8,8} & a_{8,9} \\
 & & & a_{9,4} & & a_{9,6} & a_{9,7} & & a_{9,9} \\
 & & & a_{10,4} & & a_{10,6} & a_{10,7} & & a_{10,9} \\
 & & & a_{11,4} & & a_{11,6} & a_{11,7} & & a_{11,9} \\
 & & & a_{12,4} & & a_{12,6} & a_{12,7} & & a_{12,9} \\
 & & & & a_{13,5} & a_{13,6} & & a_{13,8} & a_{13,9} \\
 & & & & a_{14,5} & a_{14,6} & & a_{14,8} & a_{14,9} \\
 & & & & a_{15,5} & a_{15,6} & & a_{15,8} & a_{15,9} \\
 & & & & a_{16,5} & a_{16,6} & & a_{16,8} & a_{16,9}
\end{bmatrix}$$

FIG. 1. *A sample matrix A.*

The elimination tree has n nodes, one for each column of A. Its structure depends on the column ordering of A and contains no information on the number of rows in A. Lacking detailed information on the rows in A, the elimination tree nonetheless provides the following constraint on the precedence of operations in a QR factorization: a row in the factor R cannot be completed until all the rows corresponding to its descendants in the elimination tree have been computed. Recursively, a row cannot be computed before any of its children.

The multifrontal algorithm turns this necessary condition into a sufficient condition for a well-organized data management scheme. A simple (nodal) multifrontal QR algorithm is given in Figure 4.

This multifrontal algorithm is superior to a simple columnwise algorithm because

$$\begin{bmatrix}
r_{11} & & r_{13} & & & & r_{17} & & r_{19} \\
 & r_{22} & r_{23} & & & & & r_{28} & r_{29} \\
 & & r_{33} & & & & r_{37} & r_{38} & r_{39} \\
 & & & r_{44} & & r_{46} & r_{47} & & r_{49} \\
 & & & & r_{55} & r_{56} & & r_{58} & r_{59} \\
 & & & & & r_{66} & r_{67} & r_{68} & r_{69} \\
 & & & & & & r_{77} & r_{78} & r_{79} \\
 & & & & & & & r_{88} & r_{89} \\
 & & & & & & & & r_{99}
\end{bmatrix}$$

FIG. 2. *A's upper triangular factor R.*

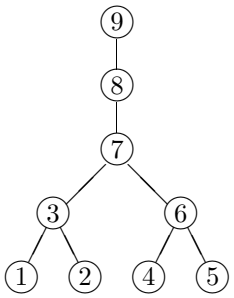


FIG. 3. *Elimination tree for R in Figure* 2.

FOR $i = 1$ **TO** n **DO**

Assemble the frontal matrix for node i, consisting of all rows of A with first nonzero in column i and all generated data rows computed by the children of node i.

Compute the QR factorization of the frontal matrix.

Save the first row of the factor as the ith row of R.

Save the remaining generated data rows for node i's parent.

END

FIG. 4. *Nodal multifrontal QR factorization.*

its row merge step includes a larger set of rows. By including more generated data rows we achieve better reduction of the operation count and we are better able to use dense vector operations. However, the major change comes in data structures. By collecting *all* generated data rows from all children at each step, all generated data rows are merged by the parent of the node at which they were computed. It follows directly that any depth-first traversal of the elimination tree can be implemented using a stack to represent all unprocessed generated data rows. Further, good algorithms are available to choose a traversal that minimizes the storage needed for the stack [29].

Some additional notation will be useful later. Denote by A_i the matrix whose

rows correspond to those rows of A that have their first nonzero in column i. We will denote the frontal matrix for the ith step by F_i. (When the ith step corresponds to a leaf node of the elimination tree, $F_i = A_i$.) At each step we compute the QR factorization of the frontal matrix as

$$F_i = Q_i \begin{bmatrix} r_{ii} & s_i \\ 0 & T_i \end{bmatrix}.$$

The factor's nonzero structure is completely determined by the rows that are used to form it and is independent of the manner of computing the QR factorization. The first row of the triangular factor of F_i is the ith row of R, with diagonal entry r_{ii} and off-diagonal entries denoted by the vector s_i.

The generated data rows from F_i are denoted by T_i. Although row merging usually produces triangular matrices T_i, it is not necessary that T_i be triangular. The matrix of generated data rows, T_i, can be trapezoidal.

If we examine the trapezoidal matrix

$$T_i = \begin{bmatrix} t^i_{11} & t^i_{12} & t^i_{13} & \cdots \\ 0 & t^i_{22} & t^i_{23} & \cdots \\ 0 & 0 & t^i_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

more closely, we discover that all column indices correspond to ancestors of node i in the elimination tree. Although only the first row of T_i contains a nonzero entry in the column corresponding to the parent of node i, all of these rows will be merged by the parent.

In the general case, the frontal matrix F_i is the result of assembling original and generated data rows. Notationally,

$$F_i = \begin{bmatrix} A_i \\ T_{i_1} \\ \vdots \\ T_{i_k} \end{bmatrix},$$

where nodes $i_1, \ldots, i_k$ are the children of node i. This represents the input to the ith factorization step. The component triangular pieces provide a structure to F_i that we exploit in the next section.

2.2. Use of dense matrix storage and kernels. The trapezoidal matrix T_i will only have nonzeros in those columns in which row i of R has off-diagonal nonzeros. Liu [28] showed that it is usually the case that any entry within these columns that can be nonzero, that lies within the trapezoidal structure, is nonzero. Therefore, we can effectively represent T_i by a dense (trapezoidal) matrix with the same number of rows as T_i and one fewer column than row i of R has nonzeros.

The frontal matrix F_i can similarly be represented by a dense rectangular matrix with as many columns as row i of R has nonzeros and as many rows as F_i. The QR factorization that comprises the row merge step can take place within this dense representation, using standard dense matrix operations. We use the structure of any generated data rows in F_i to further reduce the operation count. We rearrange the rows of F_i to form a staircase matrix, that is, we place all rows with first nonzero in

the first column first, those rows with first nonzero in the second column next and so on. Using the sample matrix problem, the reordered rows of the frontal matrix F_3 are

$$F_3 = \begin{bmatrix} t^2_{11} & 0 & t^2_{12} & t^2_{13} \\ t^1_{11} & t^1_{12} & 0 & t^1_{13} \\ 0 & t^1_{22} & 0 & t^1_{23} \\ 0 & 0 & t^2_{22} & t^2_{23} \\ 0 & 0 & 0 & t^2_{33} \\ 0 & 0 & 0 & t^1_{33} \end{bmatrix}.$$

We exploit this structure in applying Householder transformations to compute the QR factorization of F_i.

Using condensed forms of the trapezoidal and frontal matrices enables us to perform all numerical operations as dense vector operations. Indexed or indirectly addressed operations are needed only in moving data from the data rows into the dense representation of F_i. Even this sparse overhead can be reduced further through use of the natural block structure of the factor R, the topic of the next section.

2.3. Using supernodes for higher performance. Supernodes allow us to combine separate row merge steps into a single step. This will give increased speed, slightly reduce the number of operations, and reduce data movement. We define this structure by first defining a fundamental supernode.

DEFINITION 2.2 (fundamental supernode). *A fundamental supernode is a maximal set $\{i_1, i_2, \ldots, i_s\}$ of nodes in the elimination tree of A^TA, such that i_{k-1} is the only child of node i_k and such that the row structure of row i_1 of the factor R contains the row structure of row i_k, for $k = 2, \ldots, s$.*

The s factorization steps for these nodes in the nodal multifrontal factorization consists of successive QR factorizations of the frontal matrices

$$F_{i_1} = \begin{bmatrix} A_{i_1} \\ T_{j_1} \\ \vdots \\ T_{j_k} \end{bmatrix}, F_{i_2} = \begin{bmatrix} A_{i_2} \\ T_{i_1} \end{bmatrix}, F_{i_3} = \begin{bmatrix} A_{i_3} \\ T_{i_2} \end{bmatrix}, \ldots, F_{i_k} = \begin{bmatrix} A_{i_k} \\ T_{i_{k-1}} \end{bmatrix},$$

where nodes $j_1, \ldots, j_k$ are the children of node i_1. An equivalent and more efficient way to compute the same rows of R and the same generated data rows T_{i_k} is to compute the QR factorization of

$$F_I = \left[\begin{array}{c} A_{i_1} \\ A_{i_2} \\ \cdots \\ A_{i_k} \\ \hline T_{j_1} \\ \vdots \\ T_{j_k} \end{array}\right].$$

By combining the s nodal merges into a single step, we reduce data movement and we slightly reduce the operation count. We increase the efficiency of our kernels because the lengths of the vectors increase. We do fewer, longer vector operations, still on dense vectors, without increasing fill in R.

Fundamental supernodes provide a unique partitioning of the nodes in the elimination tree. Larger supernodes can be found by dropping the constraint that all but

the first node in a supernode have only a single child. There may be many partitionings into supernodes that use only the condition of common sparsity structure. Coarser partitionings enhance the benefits obtained from fundamental supernodes. A simple algorithm that finds supernodes by combining fundamental supernodes is to perform a depth-first traversal of the elimination tree. When visiting the first node of a fundamental supernode, combine this supernode with its largest numbered child supernode when the row structure of the parent supernode is a subset of the row structure of the child supernode. The frontal matrix now must incorporate the generated data rows from all children of fundamental supernodes that comprise the larger supernode.

In Figure 5 we display the steps in the supernodal multifrontal method for the model problem. In this small example there is only one nontrivial fundamental supernode, consisting of nodes 7, 8, and 9. A larger supernode is created by the inclusion of node 6.

We will use uppercase subscripts to denote the matrices involved in the supernodal factorization; that is, we will use A_I, F_I, T_I, or R_I, S_I where I now indicates the supernodal subscript set $\{i_1, i_2, \ldots, i_s\}$. The orthogonal reduction step computes

$$F_I = Q_I \begin{bmatrix} R_I & S_I \\ 0 & T_I \end{bmatrix}.$$

In general R_I will be an upper triangular matrix and S_I will be rectangular. The first s columns of F_I, corresponding to the indices of the supernode (and R_I), are called the fully assembled or internal columns of the frontal matrix. The other columns in F_I, which will make up S_I and T_I, are called external or generated data columns.

One subtle aspect of the general supernodal factorization of relevance to the condition estimation algorithm is that it is no longer necessary that all trapezoidal matrices of generated data contributing to the frontal matrix have their first nonzero in the first column of the frontal matrix. In our small example, the first row of the trapezoid from supernode 3 occurs in column 7. As a result, T_3 is not used in computing the first row of R_6 and T_3 makes no contribution to the sixth row of R. This possibility requires a certain amount of care in the implementation of the condition estimation scheme.

In summary, our basic QR factorization is computed by the supernodal sparse multifrontal algorithm. Our implementation initially applies the multiple minimum degree ordering [27] to the columns of A. The nodes of the elimination tree are then numbered in a postorder traversal to minimize the stack storage [29]. We currently minimize the stack storage based on the symbolic factorization of the Cholesky factor of the normal equations. This may not be optimal because it assumes that possibly trapezoidal matrices are triangular. We use dense matrix kernels for the computations. Overall this provides a very efficient and fast method for computing the QR factorization of large sparse problems [26].

In developing a sparse $RRQR$ factorization we tried to preserve the benefits of the multifrontal algorithm, yet allow pivoting. To do this it is necessary to answer two questions: how to determine if a column should be pivoted and how to pivot a column. We sought answers that fit into the multifrontal scheme, which we discuss in the next two sections. We begin by answering the question of how to detect when a column needs to be pivoted by using a condition number estimate that fits naturally into a multifrontal implementation.

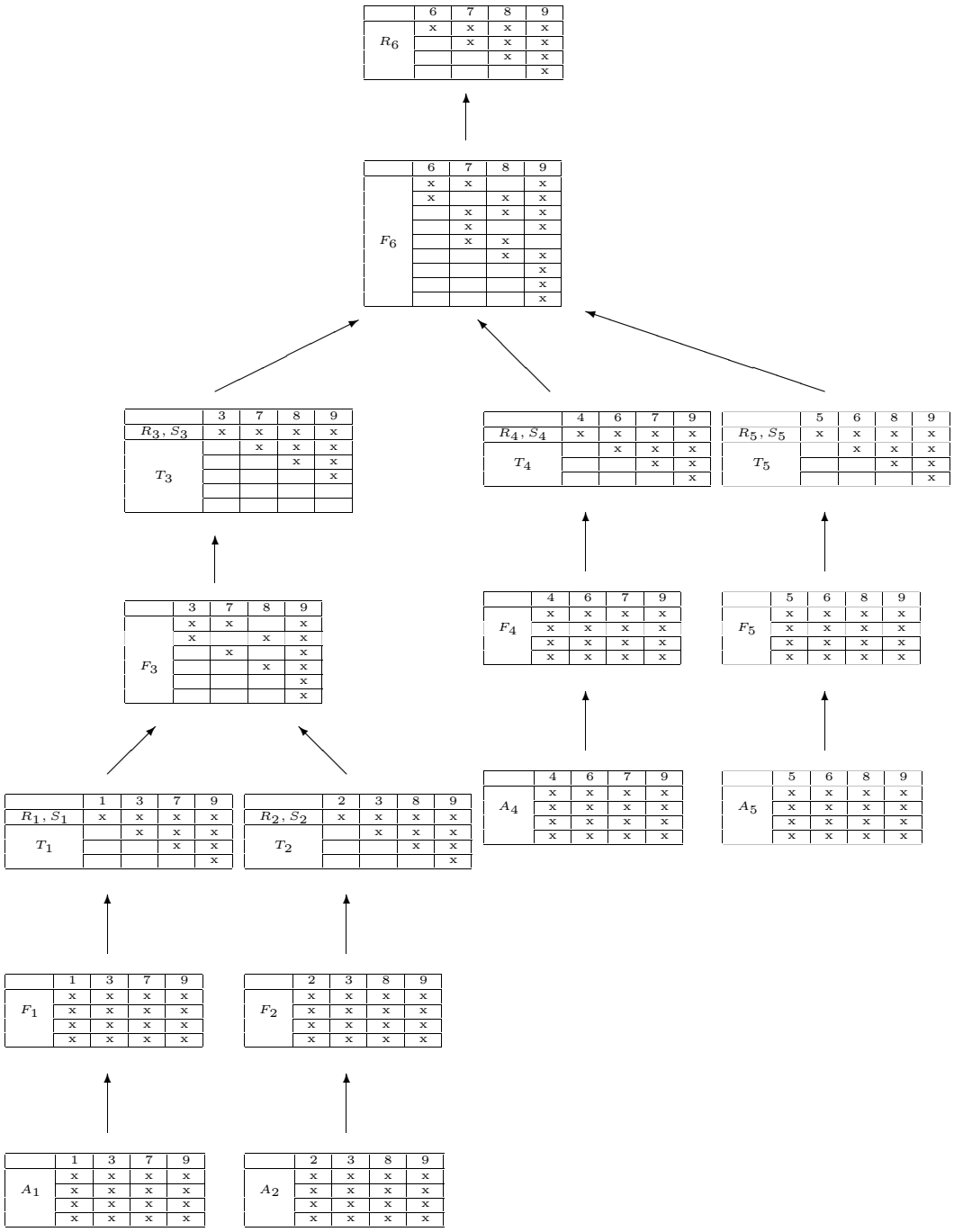


FIG. 5. *Supernodal multifrontal QR for the model problem.*

3. Condition estimation. Bischof [2] developed a method, ICE, that incrementally estimates the condition number of a triangular matrix; that is, it estimates the condition of a triangular matrix and also of all leading principal minors. This method was generalized to sparse matrices by Bischof and the present authors in [4]. The essential characteristic of this algorithm, known as SPICE, is that it uses the elimination tree structure from R in exactly the same way as the QR factorization itself. That is, the estimate of the least singular value and vector for the partial factorization up to and including a given node or supernode requires knowledge only of the partial factorization and the least singular value and vector estimates for each of the children of that node or supernode. Recursively the method computes singular value and vector estimates for submatrices corresponding to subtrees of the elimination tree.

To illustrate SPICE, assume that the current supernode has two children supernodes. For convenience we number the current supernode as 3 and its children as 1 and 2. The orthogonal reduction step for each child resulted in factorizations

$$\begin{bmatrix} R_1 & S_1 \\ 0 & T_1 \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} R_2 & S_2 \\ 0 & T_2 \end{bmatrix}.$$

Here R_I is triangular, (R_I, S_I) are the rows of the factor R with row indices in supernode I), and T_I is the trapezoidal (or triangular) matrix of generated data rows. In detail consider the columns of S_I, denoted by

$$S_I = (s_1^I, s_2^I, \ldots, s_{p_I}^I). \tag{2}$$

We also assume that we have estimates for the least singular value and vector for each of the children; that is, a vector pair (x_I, b_I) has been formed for each of the upper triangular matrices R_I such that $R_I^T x_I = b_I$, $\|b_I\|_2 = 1$ and $1/\|x_I\|_2 \approx \sigma_{\min}(R_I)$.

The assembled frontal matrix for the current supernode, F_3 is given by

$$F_3 = \begin{bmatrix} A_3 \\ T_2 \\ T_1 \end{bmatrix}.$$

The first reduction step, applying the first Householder transformation H_1, results in the matrix

$$H_1^T F_3 = \begin{bmatrix} \delta & s \\ 0 & \widetilde{T} \end{bmatrix}.$$

Hence the current partial factor is

$$\widetilde{R} = \begin{bmatrix} R_1 & 0 & s_1^1 \\ & R_2 & s_1^2 \\ & & \delta \end{bmatrix}.$$

We assume, for the moment, that s_1^1 and s_1^2 are both not equal to zero. Then SPICE finds a new estimate for the least singular value and vector by solving the optimization problem

$$\max_{\|\widetilde{b}\|_2 = 1} \|\widetilde{x}\|_2,$$

where

$$\widetilde{R}^T \widetilde{x} = \widetilde{b} \quad \text{with} \quad \widetilde{x} = \begin{bmatrix} \alpha_1 x_1 \\ \alpha_2 x_2 \\ \beta \end{bmatrix} \quad \text{and} \quad \widetilde{b} = \begin{bmatrix} \alpha_1 b_1 \\ \alpha_2 b_2 \\ \gamma \end{bmatrix}$$

for scalars α_1, α_2, and β. This only requires the inner products $c_{11} = x_1^T s_1^1$, $c_{21} = x_2^T s_1^2$, $\eta_1 = x_1^T x_1$, $\eta_2 = x_2^T x_2$, and the computation of the largest eigenvalue $\lambda_{\max}$ and the corresponding eigenvector of the 3×3 symmetric matrix

$$G = \begin{bmatrix} \eta_1 + c_{11}^2/\delta^2 & c_{11}c_{21}/\delta^2 & -c_{11}/\delta^2 \\ c_{21}c_{11}/\delta^2 & \eta_2 + c_{21}^2/\delta^2 & -c_{21}/\delta^2 \\ -c_{11}/\delta^2 & -c_{21}/\delta^2 & 1/\delta^2 \end{bmatrix} \tag{3}$$

$$= \begin{bmatrix} x_1^T x_1 & 0 & 0 \\ 0 & x_2^T x_2 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{1}{\delta^2} \begin{bmatrix} c_{11} \\ c_{21} \\ -1 \end{bmatrix} \begin{bmatrix} c_{11} & c_{21} & -1 \end{bmatrix}. \tag{4}$$

An obvious generalization to a $(k+1) \times (k+1)$ eigenproblem allows SPICE to treat cases with $k > 2$ children. The obvious simplification for a single child is identical to ICE. We fully exploit the fact that the matrices involved in the eigenvalue problems are rank one modifications of diagonal matrices. Further details are given in [4], [5], [7], [36]. Because of its recursive nature the approximate singular vector (the SPICE vector) is never required explicitly. Moreover, the SPICE algorithm in [4] never requires explicit access to the columns of the global matrix [4]. This multifrontal SPICE algorithm for fundamental supernodes is given in Figure 6. The algorithm for general supernode partitionings differs only in that the number of children at steps two to d may be greater than one, and hence the eigenproblem may be larger than 2×2.

The rank-revealing QR factorization requires the SPICE vector x in order to determine which column should be removed from R at a given step. This requires an additional n vector of storage beyond the basic SPICE algorithm's requirements. For convenience we save x in a partitioned form, with the entries corresponding to a given supernode stored contiguously with the factor data for that supernode. Also for convenience, we store the scaling factor used for each child rather than applying the scale factors directly to x. This requires one extra datum for each supernode, again stored with the corresponding factor data. This datum enables us to reconstruct the x vector whenever it is required. For further details see [33].

With this condition estimation tool in hand, we now describe how to select a column to pivot and how to carry out the pivoting operation. This will correspond to Phase I of the method.

4. Phase I: Initial QR factorization with column removal. In the multifrontal QR algorithm without pivoting there are two types of columns in a frontal matrix. The leading s columns for a supernode of s nodes correspond to the rows of R that will be computed at this factorization step. These are the fully assembled or interior columns of the supernode. The remaining columns correspond to a subset of the ancestors of the current node. These exterior columns also correspond in index to the generated data rows produced at this step.

Pivoting to reveal the rank introduces a third category, the "redundant" columns, fully assembled columns that are identified by the condition estimator as being nearly linear combinations of other columns in the partial factorization. The triangular structure of the $RRQR$ factorization (1) is achieved by reordering the columns into the new order $(1, 2, \ldots, k-1, k+1, \ldots, n, k)$ when column k is identified as redundant. In the first phase redundant columns never reenter the set of columns defining R. Thus this column, and all other redundant columns, will be fully processed only when the root supernode is factored. The effect on the sparse factorization is twofold. In the supernode I in which column k is an interior column, column k becomes the last

FOR $I = 1$ **TO** *number_of_fundamental_supernodes* **DO**

Assume supernode I has t children, m columns and d nodes;
Initialize a $t \times m$ matrix C and a $1 \times m$ vector $\hat{c}$ to zero.

FOR $i = 1$ **TO** t **DO**
Let K_i denote the ith child's index.
Pop $x_{K_i}^T s_j^{K_i}$ values from stack and scatter into ith row of C.
Pop $x_{K_i}^T x_{K_i}$ from stack and place in η_i
END

Perform SPICE on first column of supernode I

Build the $(k+1) \times (k+1)$ SPICE matrix and solve the maximum eigenvalue problem for the eigenvector $[\alpha_1, \alpha_2, \ldots, \alpha_k, \beta]^T$.
Update and save the inner product values
$\hat{c}_j = \sum_{i=1}^{k} \alpha_i c_{ij},\ j = 2, \ldots, m$
$\hat{x} = \beta, \qquad \hat{\eta} = \lambda_{\max}\ (= \sum_{i=1}^{k} \eta_i^2 \alpha_i^2 + \beta^2)$

Perform SPICE on remaining columns of supernode I

FOR $p = 2$ **TO** d **DO**
Assume new column of frontal factor is $(\hat{s},\ \delta)^T$.
Compute the global inner product $\zeta = \hat{c}_p + \hat{x}^T \hat{s}$
Solve the 2×2 SPICE eigenproblem for α and β.
Update the SPICE data:
$\hat{c}_j \leftarrow \alpha \hat{c}_j, \qquad j = p+1, \ldots, m$
$\hat{x} \leftarrow \begin{bmatrix} \alpha \hat{x} \\ \beta \end{bmatrix}, \qquad \hat{\eta} \leftarrow \lambda_{\max}\ (= \alpha^2 \hat{\eta} + \beta^2)$
END

$\hat{c}_j \leftarrow \hat{c}_j + x^T s_j^I\ (j = d+1, \ldots, m)$, where s_j^I is as in (2).
Place $\hat{c}_j, (j = d+1, \ldots, m)$ and $\hat{\eta}$ onto the stack.

END

FIG. 6. *Fundamental supernodal SParse Incremental Condition Estimation.*

numbered exterior column. The amount of storage needed for R decreases, but the storage for the generated data rows increases. As a redundant column, column k also becomes effectively a new exterior column for each ancestor of supernode I. The storage for each ancestor supernode must therefore be increased. Were we computing strictly a Type I algorithm, the redundant columns could be discarded and dynamic storage would not be a major issue. However, the redundant columns are required for Phase II and must be saved.

In the remainder of this section we describe how we determine redundant columns and how we restructure the factorization.

As we compute the factorization, we use SPICE to incrementally approximate the smallest singular value and corresponding left singular vector. We use the largest Euclidean norm of a column of A as an approximation to the largest singular value of A. The ratio of these two values provides an estimate of the condition number of

the partial factor. (SPICE could be used to approximate the largest singular value, but we currently believe that the additional accuracy is unnecessary.) The condition estimate exceeding a tolerance indicates that the submatrix for the subtree rooted at the current supernode is rank deficient.

We remove a column from this submatrix as soon as the rank deficiency is detected. We construct an approximate left singular vector x for the smallest singular value of the triangular matrix rooted at the current node, using the additional data of local $\hat{x}$ vectors and SPICE scaling factors. We then perform a backsolve on this vector with the upper triangular submatrix to form an approximate right singular vector for the smallest singular value.

The index of the component of largest absolute value in the approximate right singular vector designates the column k to be removed from the matrix. This is identical to the method proposed in Golub, Klema and Stewart [21], except that we use an approximate singular vector as in Foster [15]. The redundant column k occurs in supernode I, either the current supernode or one of its descendants. We apply the permutation $(1, 2, \ldots, k-1, k+1, \ldots, r, k)$ symmetrically to the rows and columns of the $r \times r$ partial factor for supernode I, yielding the matrix

$$\begin{bmatrix} \overline{R} & u \\ t_1 & \omega \end{bmatrix},$$

where $\overline{R}$ is the factor with the kth column u and the kth row $t = (t_1,\ \omega)$ removed. We can apply Givens rotations to this matrix to restore the triangular structure.

We exploit the sparsity of the factor by applying rotations only to the supernodes on the path from supernode I to the current supernode. We must actually permute rows and columns in the supernode I in which column k was an internal column. For the other supernodes on the path from supernode I to the root of the tree, the permutation of columns means that the existing factor data are augmented by a new row and new external column. In each case the factor data are restored to triangular form for each supernode. When the current supernode's structure has been updated, the factorization can continue. However, the redundant column will now be effectively a new exterior column for each ancestor of the current supernode. Note that in applying the rotations to t we simultaneously update the condition estimator.

There are two motivations for removing the column as soon as possible rather than after the factorization has been completed. First, the backsolve to compute the approximate right singular vector is performed only on the subtree, which reduces operations and access to the elements of the factor matrix. Otherwise the backsolve would take place on the entire matrix. Second, Givens rotations are used only to restore the partial factorization to triangular form. The remaining work associated with redundant columns at ancestor supernodes takes place as part of the standard merge process, which uses Householder transformations. Thus, fewer and faster operations are required.

Figure 7 shows the effect of restructuring the factor R if column 2 were labeled redundant during the reduction of the root supernode of our model problem. Note that row S_3 and all rows of F_6, S_6 gain a new column and a new row T_6 is added to the root of the factor. This example also shows that it is possible for supernodes to disappear entirely from the factor tree, in this case because the redundant column was the only interior column in the second supernode.

5. A bound on $\|T\|_2$. Our goal is a factorization of the form in (1), where R is reasonably conditioned and T is small in norm. Phase I of our algorithm produces

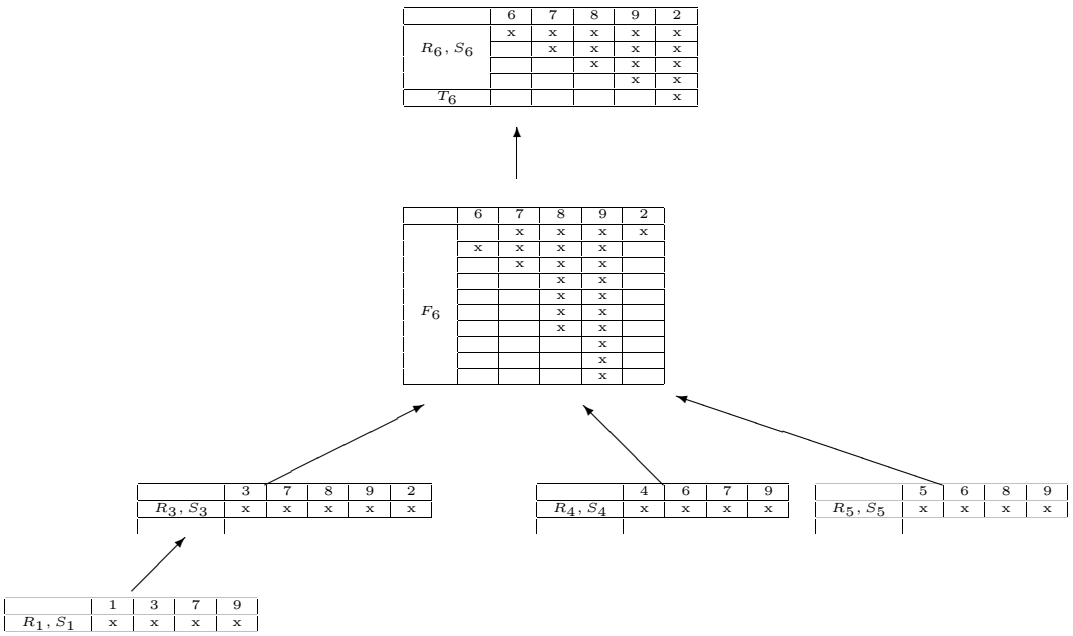


FIG. 7. *Modified factor after pivoting column* 2.

an acceptably well-conditioned R. In this section we show that $\|T\|_2$ has a bound that depends exponentially on the number of columns in T. Our analysis of $\|T\|_2$ is divided into two cases, distinguished by the presence or absence of a zero diagonal element in the factor R. It is necessary to divide the analysis because the condition estimator SPICE will fail in the presence of a zero diagonal element. We first consider the full structural rank case where R has no zero diagonal elements. In this case our analysis is almost identical to Chan [8], except that we require only approximations to the singular vectors. We only establish enough in the second case, with zero diagonal entries, to reduce it to the first case. In both cases the result of the analysis is a bound on $\|T\|_2$ similar to that in [8], where exact singular vectors were required.

We assume that we have computed a triangular factor R for the first r columns of A. We denote this submatrix as $A_{1:r}$, so that we have

$$A_{1:r} = QR.$$

Further assume that the condition estimate has exceeded a tolerance of $1/\tau$, signaling an unacceptable linear dependence. However, the last computed diagonal element of R is not zero; the matrix R appears to be of full rank. The singular value estimate from SPICE is a vector pair $(x,\ b)$ such that

$$R^T x = b,$$

where

$$\frac{1}{\|x\|_2} = \sigma_e \geq \sigma_{\min}(R) \qquad \text{and} \qquad \|b\|_2 = 1.$$

The approximation σ_e is a fairly good estimate for $\sigma_{\min}$ in the sense that $\sigma_e < \tau\sigma_{\max} \ll \sigma_{\max}$. We compute an approximate right singular vector y by solving

$$Ry = \sigma_e x$$

and determine the entry of largest magnitude y_i using the largest index in case of ties.

Let P be a permutation matrix that permutes columns $(1, 2, \ldots, r)$ into the new order $(1, 2, \ldots, i-1, i+1, \ldots, r, i)$. Let $\tilde{y} = Py$. Then

$$\|y\|_\infty = \|Py\|_\infty = |\tilde{y}_r| > 0.$$

Applying the permutation P to R produces

$$RP^T = \begin{bmatrix} R_{11} & c & S \\ 0 & \delta & w^T \\ 0 & 0 & R_{22} \end{bmatrix} P^T = \begin{bmatrix} R_{11} & S & c \\ 0 & w^T & \delta \\ 0 & R_{22} & 0 \end{bmatrix},$$

where δ is the ith diagonal element and R_{11} and R_{22} are upper triangular matrices. We now seek to bound the resulting (r, r) element of the matrix RP^T after computing the orthogonal factorization of the permuted matrix,

$$\widetilde{R} = \widetilde{Q}RP^T = \begin{bmatrix} R_{11} & R_{12} & s_1 \\ 0 & \widetilde{R}_{22} & s_2 \\ 0 & 0 & \widetilde{\delta} \end{bmatrix}.$$

Now note that

$$\sigma_e = \|\sigma_e x\|_2 = \|Ry\|_2 = \|QRP^T Py\|_2 = \|\widetilde{R}\tilde{y}\|_2 \geq |\widetilde{\delta}\tilde{y}_r|.$$

Thus,

$$|\widetilde{\delta}| \leq \frac{\sigma_e}{|\tilde{y}_r|}.$$

We now show that we can always find a satisfactory $\tilde{y}_r$ by demonstrating, with minimal assumptions on the accuracy of σ_e, that $\|y\|_2$ will be at least $1/2$. As a result the largest entry y_i satisfies $|y_i| \geq \frac{1}{2\sqrt{r}}$. Let

$$R = V\Sigma U^T$$

be the singular value decomposition of R, with the singular values ordered $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. If

$$x = \sum_{i=1}^{r} \rho_i v_i,$$

then

$$1 = \|x\|_2^2 = \sum_{i=1}^{r} \rho_i^2$$

and

$$\|b\|_2^2 = \|R^T x\|_2^2 = \|U\Sigma V^T x\|_2^2 = \sum_{i=1}^{r} \rho_i^2 \sigma_i^2 = \sigma_e^2.$$

We expect the coefficients ρ_i to be larger for larger i. By assumption, $\sigma_{\max} \gg \sigma_e$, so a $\beta > 0$ and an integer k exist such that $\sigma_k \geq (1+\beta)\sigma_e \geq \sigma_{k+1}$. Then

$$\sigma_e^2 \geq \sum_{i=1}^{k} \rho_i^2 \sigma_i^2 \geq \sum_{i=1}^{k} \rho_i^2 \sigma_k^2 \geq (1+\beta)^2 \sigma_e^2 \sum_{i=1}^{k} \rho_i^2.$$

Therefore, by dividing both sides of the inequality above by σ_e^2 we have

$$1 \geq (1+\beta)^2 \sum_{i=1}^{k} \rho_i^2 \Rightarrow \sum_{i=k+1}^{r} \rho_i^2 > 1 - \frac{1}{(1+\beta)^2} = \frac{\beta(\beta+2)}{(1+\beta)^2},$$

and for all $j > k$,

$$\sigma_j \leq (1+\beta)\sigma_e \Rightarrow \frac{1}{1+\beta} \leq \frac{\sigma_e}{\sigma_j}.$$

It follows that

$$\begin{aligned}
\|y\|_2^2 &= \sum_{i=1}^{r} \left(\frac{\sigma_e}{\sigma_i}\rho_i\right)^2 \\
&\geq \sum_{i=k+1}^{r} \left(\frac{\sigma_e}{\sigma_i}\rho_i\right)^2 \\
&\geq \left(\frac{1}{1+\beta}\right)^2 \sum_{i=k+1}^{r} \rho_i^2 \\
&\geq \left(\frac{1}{1+\beta}\right)^2 \frac{\beta(\beta+2)}{(1+\beta)^2} \\
&= \frac{\beta(\beta+2)}{(1+\beta)^4}.
\end{aligned}$$

We summarize these results in the theorem below.

THEOREM 5.1. *If the pivoting criterion of Golub, Klema, and Stewart* [21] *is used with an approximate least singular value, σ_e, and approximate right singular vector y and if there is a gap in the singular values such that there exists a $\beta > 0$ and an integer k with $\sigma_k \geq (1+\beta)\sigma_e \geq \sigma_{k+1}$, then*

$$\|y\|_2 > \frac{\beta(\beta+2)}{(1+\beta)^4}. \qquad \square$$

The value for β that gives the largest lower bound for $\|y\|_2$ is found by optimizing the function

$$f(x) = \frac{x^2+2x}{(1+x)^4}.$$

The maximum on the interval $[0,\infty)$ is achieved when $x = \sqrt{2}-1$, with $f(x) = 1/4$. It follows that $\|y\|_2 \geq 1/2$ whenever an integer k can be found such that $\sigma_k \geq \sqrt{2}\sigma_e \geq \sigma_{k+1}$. Such a k would not exist only if $\sigma_{\max} < \sigma_e$, which is unreasonable. Therefore, $\|y\|_2 > 1/2$ and so the largest entry, y_i, which indicates the redundant column, satisfies $|y_i| > \frac{1}{2\sqrt{r}}$. Thus we have chosen to remove a column so that

$$|\widetilde{\delta}| \leq 2\sqrt{r}\sigma_e.$$

Now consider the case of a structural rank deficiency in R. In this case R has a zero diagonal entry, say the kth, and the SPICE algorithm would break down with a division by zero. If we label the kth column as redundant, we may later get appreciable growth in the matrix T. This occurs because we have no lower bound on the size of the kth element of the right singular vector. In fact, in practice it is often very small. If we later label as redundant a column i, where $i < k$, the resulting matrix T can have very large off-diagonal components.

There is a simple solution to the problem—the exact nullspace vector can be computed directly. If the r by r factor is of the form

$$\widetilde{R} = \begin{bmatrix} R & u \\ 0 & 0 \end{bmatrix},$$

with u a column vector, the right singular vector corresponding to the singular value 0 is the vector y given by

$$z = \begin{bmatrix} R^{-1}u \\ -1 \end{bmatrix},$$
$$y = \frac{z}{\|z\|_2}.$$

The cost of computing this vector is actually slightly less than the SPICE vector x, because we must reconstruct the scalings of the latter. Since $\widetilde{R}y = 0$ and $\|y\|_2 = 1$, there exists a component y_i of y such that $|y_i| > \frac{1}{\sqrt{2r}}$ and thus we will satisfy the conclusion of Theorem 5.1 for the structurally rank-deficient case as well.

Once a column has been pivoted out of the factor, it is no longer used in condition estimation. This means that if we were to augment the computed y vectors with zeros so that they would all be of length n, then the $\widetilde{y}$ vectors would be zero past their largest component. This fact, coupled with the bound in Theorem 5.1, enables us to mimic the proofs in Chan [8]. We substitute $\frac{1}{2\sqrt{j}}$ for the lower bound of the singular vector and use τ, the pivot threshold, for the upper bound of the computed σ_e's to show that even for our left to right factorization the following theorem holds.

THEOREM 5.2. *With the hypothesis of Theorem* 5.1*, the triangular matrix* T *constructed obeys the bounds*

$$|t_{ij}| \leq 2^{j-i+1}\tau\sqrt{n}. \qquad \square$$

As a result the bound on $\|T\|_2$ is of $O(2^{k+1})$, where k is the order of T.

6. Phase II: Keeping $\|T\|_2$ small. In the previous section we established a bound on $\|T\|_2$, which depends exponentially on the order of T. This bound could become unacceptably large when a large number of columns are redundant or if the heuristic condition estimator has performed poorly. Some of our applications require computing nullvectors for the operator, where it is imperative that any approximate nullvectors be good approximations. This motivates our second phase, which guarantees that T has a small norm.

After we have computed the factorization

$$\begin{bmatrix} R & S \\ 0 & T \end{bmatrix},$$

we use the bounds in Hong and Pan [24] to determine if there is a need to reinstate redundant columns into R. The second phase removes columns from $\begin{bmatrix} S^T & T^T \end{bmatrix}^T$

and places the columns back into the factor R whenever

$$\|T\|_F > \sqrt{(n-k)(k+1)}\tau.$$

When needed, we perform a dense $RRQR$ factorization of T, using Golub's column interchange criterion [20], which orders the columns of the reduced matrix by norm. This yields a new triangular factor $\widehat{T}$. The column interchanges are also applied to the columns of S, producing $\widehat{S}$. The orthogonal factor of A is then partitioned as

$$\left[\begin{array}{cc|c} R & \widehat{S}_1 & \widehat{S}_2 \\ 0 & \widehat{T}_{11} & \widehat{T}_{12} \\ \hline 0 & 0 & \widehat{T}_{22} \end{array}\right],$$

where $\widehat{T}_{22}$ is the largest trailing principal submatrix of $\widehat{T}$ with Frobenius norm less than $((n-k)(k+1))^{1/2}\,\tau$. In the case where A^TA is reducible, this process would be done for each tree in the resulting elimination forest.

Phase II is consistent with the multifrontal approach in that all of the numeric operations take place in the final dense block. It has the effect of introducing much less sparse columns in the final triangular factor, since the reinstated columns contain fill from all supernodes on the path from their original interior position to the root. However, no additional storage is needed for the data in R and S.

This would not be the case if we allowed more general interchanges between R and the set of redundant columns. Were we at this stage to attempt to remove any column from R, correcting the triangular structure would produce fill in the newly redundant column corresponding to all entries in the columns of $[\ \widehat{S_1}\quad \widehat{T}_{11}\ ]^T$. One effect of Phase II is then that the condition number of the final triangular factor

$$\widehat{R} = \left[\begin{array}{cc} R & S_1 \\ 0 & \widehat{T}_{11} \end{array}\right]$$

can be larger than that of R. We do not attempt to correct this because of the cost involved.

7. Test results. In this section we present some preliminary results on the numerical performance of the method. The test matrices are matrices that arise in the triangulation of data by cubic b-splines; see [22] for a detailed explanation and example. In Table 1 we present the matrices, dimensions, numerical rank, and the gap that defines the numerical rank. All of these test problems are well-posed in the sense of having a well-defined gap.

TABLE 1
Test matrices and their gap.

matrix	rows	columns	% nonzeros	rank	σ_{rank}	$\sigma_{\text{rank+1}}$
triang01	130	49	32	48	8.4E−8	2.2E−18
scatdt06	72	50	4.7	46	3.4E−1	7.8E−16
scatdt12	101	92	17	46	7.5E−3	1.3E−14
triang02	130	100	16	79	2.3E−9	3.7E−16
triang04	9100	1156	1.4	1022	2.1E−8	7.4E−16

In all cases our method was able to detect the numerical rank and produce a well-conditioned matrix R and a matrix T of small norm. We present these results in Table 2, listing the actual and estimated smallest singular value of R as well as

$\|T\|_F$ and $k = n - r$, the order of T. The matrix triang02 has many structural rank deficiencies. If columns with zero diagonal entries are always removed rather than those labeled by constructing the singular vector, it becomes necessary to reinstate a column of T into the factor R in the second phase of the algorithm. This is due to additional columns being removed. If these additional columns had been pivoted before the zero diagonal columns, the diagonal element would not have been zero at all! For example, consider the matrix

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

If column 2 is pivoted to the end of the matrix and then column 1, we have a trailing matrix T of norm ≈ 1. On the other hand if column 1 had been pivoted first, there is no longer a zero pivot.

TABLE 2
Numerical accuracy ($\tau = 3.3\text{E}-9$).

matrix	$\sigma_{\min}(R)$	$\tilde{\sigma}_{\min}(R)$	$\|T\|_F$	k
triang01	8.4E−8	2.8E−7	8.6E−22	1
scatdt06	3.3E−1	4.8E−1	5.1E−16	4
scatdt12	5.5E−3	6.2E−3	4.8E−14	46
triang02	2.3E−9	4.1E−9	1.1E−18	21
triang04	1.9E−8	2.5E−8	2.8E−16	134

Table 3 lists the costs of different parts of the algorithm as percentages of total operations. Costs are given for computing and applying the orthogonal transformations (% factor), solving the seminormal equations (% solve), estimating the condition number (% condition), determining which column to pivot, and annihilating the removed row through to the current supernode (% pivot). All floating point operations and square roots are counted.

TABLE 3
Numerical cost as a percentage of total factorization cost.

matrix	% factor	% solve	% condition	% pivot
triang01	93.	3.2	3.3	0.14
scatdt06	60.	7.4	31.	5.1
scatdt12	96.	.90	2.2	1.1
triang02	83.	4.6	9.4	2.9
triang04	98.	.67	.65	1.0

The cost of determining the rank with our method is relatively inexpensive, even for problem scatdt12 where the nullity is half of the matrix.

8. Future work. In this section we summarize features we would like to see incorporated into later versions of the algorithm and additional topics to be explored.

The method of column reinstatement should incorporate the condition estimation information. As noted, this requires saving the inner product values for postponed columns, but it will result in an more robust scheme by enabling us to estimate the condition of the final matrix R.

Bischof and Hansen [3] have an alternative selection criterion that predicts the bound on $\|T\|_2$ rather than just a bound on the new diagonal element of T. This requires saving the computed approximate right singular vectors. It requires $O(kn)$

additional work, where k is the nullity, and only an additional n vector of storage. Alternative selection criteria should also be explored to identify the most offending column in terms of satisfying both Problem I and Problem II simultaneously.

It would be valuable to know when a problem is not well posed in the sense that there is no significant gap in the singular values. Numerical rank is ill defined in such a case. One could estimate the two smallest singular values of R to see if the numerical rank of the partial factor is well defined. Alternatively the current algorithm could save the largest of the singular value estimates that occur at steps that require a pivot. Since we also have the condition of the matrix R on completion, these two values would approximate the gap of the singular values. (Estimating the two smallest singular values has the added advantage of increasing the accuracy of the SPICE estimate for the smallest singular value.)

Dense rows which are to be held out of the matrix and then later incorporated implicitly by means of the Sherman, Morrison, and Woodbury formula were addressed in a quite straightforward manner in [25]. This requires the use of an adaptive condition estimation method such as ACE [34] to estimate the condition number of the final matrix.

The effectiveness of the incorporation of iterative refinement for rank-deficient least squares problems must be determined. The method of iterative refinement for rank-deficient least squares problems was investigated in [35], but only in describing how the method can be applied.

Finally, parallel algorithms or parallel variants of this algorithm should be developed, as the effectiveness of current algorithms leads to applications with much larger problem sizes and computational demands. These demands are already very near. We will need alternative parallel methods that can match or exceed the numerical reliability of our current formulation.

Acknowledgments. The authors thank Cleve Ashcraft for his timely reading of this paper and his many helpful suggestions, as well as for Figure 5 in the text.

REFERENCES

[1] J. L. Barlow and U. B. Vemulapati, *Rank detection methods for sparse matrices*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1279–1297.

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

[3] C. H. Bischof and P. C. Hansen, *A block algorithm for computing rank-revealing QR factorizations*, Numer. Algorithms, 2 (1992), pp. 371–392.

[4] C. H. Bischof, J. G. Lewis, and D. J. Pierce, *Incremental condition estimation for sparse matrices*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 644–659.

[5] C. H. Bischof and P. T. P. Tang, *Robust Incremental Condition Estimation*, Technical Report, MCS-P225-0391, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 1991.

[6] C. H. Bischof and C. Van Loan, *The WY representation for products of Householder matrices*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. s2–s13.

[7] J. R. Bunch, C. R. Nielsen, and D. C. Sorensen, *Rank-one modification of the symmetric eigenproblem*, Numer. Math., 31 (1978), pp. 31–48.

[8] T. F. Chan, *Rank revealing QR factorizations*, Linear Algebra Appl., 88/89 (1987), pp. 67–82.

[9] S. Chandrasekaran and I. Ipsen, *On Rank Revealing QR Factorizations*, Research Report, YALEU/DCS/RR-880, Department of Computer Science, Yale University, New Haven, CT, December 1991.

[10] T. F. Chan and P. C. Hansen, *Some applications of the rank revealing QR factorization*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 724–741.

[11] E. C.-H. L. Chu, *Orthogonal Decomposition of Dense and Sparse Matrices on Multiprocessors*, Ph.D. thesis, Technical Report CS-88-08, University of Waterloo, Waterloo, Ontario, Canada, March 1988.

[12] 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.

[13] I. S. Duff and J. K. Reid, *The multifrontal solution of unsymmetric sets of linear equations*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 633–641.

[14] D. R. Ferguson and T. A. Grandine, *On the Construction of Surfaces Interpolating Curves: 1. Gordon Surfaces Revisited*, Technical Report SCA-TR-105, Boeing Computer Services, Seattle, WA, 1988.

[15] L. V. Foster, *Rank and null space calculations using matrix decompositions without column interchanges*, Linear Algebra Appl., 74 (1986), pp. 47–71.

[16] J. A. George and M. T. Heath, *Solution of sparse linear least squares problems using Givens rotations*, Linear Algebra Appl., 34 (1980), pp. 69–83.

[17] J. A. George and J. W. H. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall Inc., Englewood Cliffs, NJ, 1981.

[18] J. A. George and J. W. H. Liu, *Householder reflections versus Givens rotations in sparse orthogonal decompositions*, Linear Algebra Appl., 88/89 (1987), pp. 223–238.

[19] J. A. George and E. Ng, *SPARSPAK: Waterloo Sparse Matrix Package User's Guide for SPARSPAK-B*, Research Report No. CS-84-37, University of Waterloo, Dept. of Computer Science, Waterloo, Ontario, Canada, 1984.

[20] G. H. Golub, *Numerical methods for solving linear least squares problems*, Numer. Math., 7 (1965), pp. 206–216.

[21] G. H. Golub, V. Klema, and G. W. Stewart, *Rank Degeneracy and Least Squares Problems*, Technical Report TR-456, Dept. of Computer Science, University of Maryland, Baltimore, MD, 1976.

[22] T. A. Grandine, *Rank Deficient Interpolation and Optimal Design: An Example*, Technical Report SCA-TR-113, Boeing Computer Services, Seattle, WA, 1988.

[23] M. T. Heath, *Some extensions of an algorithm for sparse linear least squares problems*, SIAM J. Sci. Statist. Comput., 3 (1982), pp. 223–237.

[24] P. Hong and C.-T. Pan, *The rank revealing QR decomposition and SVD*, Math. Comp., 58 (1992), pp. 213–232.

[25] J. G. Lewis and D. J. Pierce, *A robust sparse RRQR factorization*, Presentation at the annual SIAM conference in Los Angeles, CA, July 1992.

[26] J. G. Lewis, D. J. Pierce, and D. K. Wah, *Multifrontal Householder QR factorization*, Technical Report ECA-TR-127, Boeing Computer Services, Seattle, WA, 1989.

[27] J. W. H. Liu, *Modification of the minimum degree algorithm by multiple elimination*, ACM Trans. Math. Software, 11 (1985), pp. 141–153.

[28] J. W. H. Liu, *On general row merging schemes for sparse givens transformations*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1190–1211.

[29] J. W. H. Liu, *Equivalent sparse matrix reordering by elimination tree rotations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 424–444.

[30] J. W. H. Liu, *The role of elimination trees in sparse factorizations*, SIAM J. Matrix Anal. Appl., 12 (1990), pp. 134–172.

[31] P. Matstoms, *The Multifrontal Solution of Sparse Linear Least Squares Problems*, Thesis, Department of Mathematics, Linköping University, Linkoping, Sweden, October 1991.

[32] C.-T. Pan and P. T. P. Tang, *Bounds on singular values revealed by QR factorizations*, in SVD and Signal Processing III, M. Moonen and B. De Moor, eds., Elsevier, New York, 1995.

[33] D. J. Pierce and J. G. Lewis, *Sparse Multifrontal Rank Revealing QR factorization*, Technical Report MEA-TR-193, Boeing Computer Services, Seattle, WA, 1992.

[34] D. J. Pierce and R. J. Plemmons, *Fast adaptive condition estimation*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 274–291.

[35] Y. C. Pierce, *Iterative Refinement of Least Squares Solutions for Rank Deficient Problems*, Technical Report MEA-TR-192, Boeing Computer Services, Seattle, WA, 1992.

[36] D. C. Sorensen and P. T. P. Tang, *On the orthogonality of eigenvectors computed by a divide-and-conquer method*, SIAM J. Numer. Anal., 28 (1991), pp. 1752–1775.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 181–190, January 1997

014

VERTICAL BLOCK HIDDEN Z-MATRICES AND THE GENERALIZED LINEAR COMPLEMENTARITY PROBLEM*

S. R. MOHAN† AND S. K. NEOGY†

Abstract. In this paper we introduce vertical block hidden Z-matrices and study their minimality and complementarity properties.

Key words. generalized linear complementarity problem, vertical block Z-matrices, vertical block hidden Z-matrices, VLCP, least element

AMS subject classification. 90C33

PII. S0895479894271147

1. Introduction.

Given a square matrix M of order n and a $q \in R^n$ the linear complementarity problem is to find $w \in R^n$ and $z \in R^n$ such that

$$w - Mz = q, \quad w \geq 0, \; z \geq 0, \tag{1.1}$$

$$w^t z = 0. \tag{1.2}$$

The linear complementarity problem is well studied in the literature. For the latest books see Cottle, Pang, and Stone [2] and Murty [18]. In [14], Lemke proposes an algorithm which either computes a solution to the linear complementarity problem or shows that there is no solution to (1.1) and (1.2). We call a square matrix $M = ((m_{ij}))$ of order n a Z-matrix or say that $M \in Z$ if $m_{ij} \leq 0$, $i \neq j$, $1 \leq i \leq n$, $1 \leq j \leq n$. The linear complementarity problem with a Z-matrix has a number of applications. See [2] and [26]. Z-matrices have a least element property related to their complementarity property which has been observed by Cottle and Veinott [3].

The generalized linear complementarity problem with a vertical block matrix of order $m \times k$ was introduced by Cottle and Dantzig [1]. Their statement of this problem is as follows: Given an $m \times k$ $(m \geq k)$ vertical block matrix M of type $(m_1, m_2, \ldots, m_k)$ and $q \in R^m$ where $m = \sum_{j=1}^{k} m_j$, find $w \in R^m$ and $z \in R^k$ such that

$$w - Mz = q, \quad w \geq 0, \; z \geq 0, \tag{1.3}$$

$$z_j \prod_{i=1}^{m_j} w_i^j = 0, \quad j = 1, 2, \ldots, k. \tag{1.4}$$

This problem is denoted as VLCP(q, M).

Cottle and Dantzig [1] extended Lemke's algorithm to solve the above problem. They have also extended some of the properties of the square P-matrix to the vertical block P-matrix.

* Received by the editors July 13, 1994; accepted for publication (in revised form) by R. Cottle February 5, 1996.
http://www.siam.org/journals/simax/18-1/27114.html
† Indian Statistical Institute, 7, S. J. S. Sansanwal Marg, New Delhi 110016, India (srm@isid.ernet.in, skn@isid.ernet.in).

VLCP, or the vertical linear complementarity problem, has not been studied extensively until recently, although Lemke [15] as early as 1970 anticipated valuable applications of this problem. Recently, a number of applications of this problem have been noted in the literature. In [6], Ebiefung and Kostreva introduce a generalized Leontief input-output linear model and formulate it as a VLCP. This model can be effectively used for the problem of choosing a new technology and also for solving problems related to energy commodity demands, international trade, multinational army personnel assignment, and pollution control. In [12], Gowda and Sznajder introduce a generalized bimatrix game and formulate a special case of this as a VLCP. A slightly more general form of the VLCP also occurs in control theory [21], [22].

There have also been other generalizations of the linear complementarity problem motivated by certain other applications. The horizontal linear complementarity problem arises in nonlinear networks. See [8], [9], and [27]. Oh [19] has formulated a mixed lubrication problem as a generalized nonlinear complementarity problem.

The VLCP has been studied by Szanc [23]. A more general version in the setting of a finite dimensional lattice gives the generalized order linear complementarity problem studied by Gowda and Sznajder [11]. However, when specialized to R^n, the generalized order linear complementarity problem is seen to be equivalent to the VLCP. See Gowda and Sznajder [11]. Generalizations of P_0- and Z-matrices have been studied by Ebiefung and Kostreva [5] and Sznajder and Gowda [25]. See also [7] and [24]. The extended generalized order linear complementarity problem was considered by Goeleven [10], Gowda and Sznajder [11], and Isac and Goeleven [13].

Mangasarian [16] while studying the classes of linear complementarity problems solvable by a single linear program introduced a class of matrices which later came to be named as the class of hidden Z-matrices in [20].

A square matrix M of order n is called a hidden Z-matrix if there exist square matrices of order n, X and Y, $X \in Z$, $Y \in Z$ such that (i) $MX = Y$ and (ii) there exist nonnegative vectors r, $s \in R^n$ such that $r^t X + s^t Y > 0$.

The class of hidden Z-matrices also possesses a least element property which is related to complementarity. For a study of this property see [2]. The least element theory for hidden Z-matrices was motivated by the observation of Mangasarian [16] that the linear complementarity problem with a hidden Z-matrix can be solved as a single linear programming problem. For related results see also [17].

Recently, Ebiefung and Kostreva [4] have studied the generalized linear complementarity problem with a vertical block Z-matrix. Complementarity and least element properties and a computational scheme using principal pivoting were studied in this paper.

The present work is motivated partly by a question which naturally arises from the work of Ebiefung and Kostreva [4] and Mangasarian [16, 17]: what is the largest class of vertical block matrices for which the associated VLCP has the least element property and hence can be solved as a single linear programming problem? This also has an implication for the class of VLCPs which has polynomial time complexity. Surprisingly, unlike in the generalization of other properties of square matrices, the required generalization of the hidden Z-property does not depend upon the representative submatrices. We introduce the class of vertical block hidden Z-matrices and study the associated minimality and complementarity properties.

In section 2, we present the required notations and definitions. In section 3, we study the least element and complementarity property possessed by vertical block hidden Z-matrices. In section 4, we present some characterization theorems for vertical

block hidden K-matrices.

2. Definitions and notation. By writing $A \in R^{m\times n}$, we denote that A is a matrix of real entries with m rows and n columns. For any matrix $A \in R^{m\times n}$, a_{ij} denotes the ith row jth column entry and $\mathrm{Pos}(A)$ denotes the nonnegative cone generated by columns of A. If $A \in R^{m\times n}$ and $J \subseteq \{1, 2, \ldots, m\}$, A_J denotes the submatrix of A consisting of the rows of A whose indices are in J. $A_{\cdot i}$ denotes the ith column and $A_{i\cdot}$, the ith row of A. If $A \in R^{m\times n}$, $J_1 \subseteq \{1, 2, \ldots, m\}$ and $J_2 \subseteq \{1, 2, \ldots, n\}$, then $A_{J_1J_2}$ denotes the submatrix of A consisting of only the rows and columns of A whose indices are in J_1 and J_2, respectively. Any vector $x \in R^n$ is a column vector unless otherwise specified. x^t denotes the transpose of x. For any two vectors $x, y \in R^n$, we define $\min(x, y)$ as the vector whose ith coordinate is $\min(x_i, y_i)$. Let M be a vertical block matrix of order $m \times k$ and type $(m_1, \ldots, m_k)$ and $q \in R^m$ be given. The set $\mathrm{FEA}(q, M) = \{(w, z) \,|\, w \in R^m,\ z \in R^k,\ (w, z) \text{ satisfies } (1.3)\}$ is called the *feasible region* of $\mathrm{VLCP}(q, M)$ and any vector in $\mathrm{FEA}(q, M)$ is called a *feasible vector.*

Let C be a convex cone. We say that C is a *pointed* convex cone if C does not contain any linear subspace except $\{0\}$. If C is a pointed convex cone in R^n, C induces a partial ordering of vectors in R^n defined as follows : $x \preceq (C)\, y$ if $y - x \in C$. We call this partial ordering the *cone ordering* induced by C. In particular, in this paper we consider the cone ordering induced by C where $C = \mathrm{Pos}(X)$ for some nonsingular X.

A matrix $M \in R^{n\times n}$ is said to be a P_0-matrix (P-matrix) if all its principal minors are nonnegative (positive). Such a matrix is called a K-matrix if it is both a Z- and a P-matrix.

DEFINITION 2.1. *Consider a rectangular matrix $M \in R^{m\times k}$ with $m \geq k$. Suppose M is partitioned row-wise into k blocks in the form*

$$M = \begin{bmatrix} M^1 \\ M^2 \\ \vdots \\ M^k \end{bmatrix},$$

where each $M^j = ((m^j_{rs})) \in R^{m_j\times k}$ with $\sum_{j=1}^k m_j = m$. Then M is called a vertical block matrix of type $(m_1, m_2, \ldots, m_k)$.

DEFINITION 2.2. *A submatrix of size k of M is called a* representative submatrix *if its jth row is drawn from the jth block M^j of M.*

Remark 2.1. If $m_j = 1,\ j = 1, \ldots, k$, then M is a square matrix. Thus, a vertical block matrix is a natural generalization of a square matrix. Clearly, a vertical block matrix of type $(m_1, m_2, \ldots, m_k)$ has at most $\prod_{j=1}^k m_j$ distinct representative submatrices.

Let $J_1 = \{1, 2, \ldots, m_1\}$ and let $J_i = \{\sum_{j=1}^{i-1} m_j + 1, \sum_{j=1}^{i-1} m_j + 2, \ldots, \sum_{j=1}^{i} m_j\}$, $2 \leq i \leq k$.

The vectors $q, w \in R^m$ in (1.3) are decomposed to conform to the partition of M into blocks of M^j, $1 \leq j \leq k$, i.e.,

$$q = \begin{bmatrix} q^1 \\ q^2 \\ \vdots \\ q^k \end{bmatrix} \text{ and } w = \begin{bmatrix} w^1 \\ w^2 \\ \vdots \\ w^k \end{bmatrix},$$

where $q^j = (q_i^j)$ and $w^j = (w_i^j)$ are $m_j \times 1$ column vectors.

DEFINITION 2.3. *A vertical block matrix M of type $(m_1, m_2, \ldots, m_k)$ is called a* vertical block Z-matrix *if all its representative submatrices are Z-matrices. Vertical block P_0 (P)-matrices are also defined in a similar manner.*

DEFINITION 2.4. *Let $M \in R^{m\times k}$ be a vertical block matrix of type $(m_1, m_2, \ldots, m_k)$. M is called a* vertical block hidden Z-matrix *if there exists a Z-matrix $X = ((x_{ij})) \in R^{k\times k}$ and a vertical block Z-matrix $Y = ((y_{ij})) \in R^{m\times k}$ of the same type as M and nonnegative vectors $r \in R^k$, $s \in R^m$ such that*

(i) $MX = Y$,

(ii) $r^t X + s^t Y > 0$.

LEMMA 2.1. *Let M be a vertical block hidden Z-matrix. Let $X \in R^{k\times k}$ be any Z-matrix and $Y \in R^{m\times k}$ be a vertical block Z-matrix of the same type as M satisfying the conditions of Definition* 2.4. *Then X is nonsingular and there exists an index set $\alpha \subseteq \{1, 2, \ldots, k\}$ such that the matrix*

$$W = \begin{bmatrix} X_{\alpha\,\alpha} & X_{\alpha\,\bar{\alpha}} \\ V_{\bar{\alpha}\,\alpha} & V_{\bar{\alpha}\,\bar{\alpha}} \end{bmatrix}$$

is in K, where V is a representative submatrix of Y corresponding to a representative submatrix G of M.

Proof. Let r, s be as in Definition 2.4. Let $p = X^t r + Y^t s > 0$. Hence, $Ax = p$ where $A = [X^t, Y^t] \in R^{k\times(m+k)}$, $x \geq 0$ has a solution $x = \begin{bmatrix} r \\ s \end{bmatrix}$.

We now proceed as in the proof of Theorem 3.11.17 of Cottle, Pang, and Stone [2, p. 207] to conclude the proof of the lemma. □

We also observe the following result.

PROPOSITION 2.1. *Let M be a vertical block hidden Z-matrix with X and Y as any matrices satisfying the conditions of Definition* 2.4. *Then there is at least one representative submatrix of M which is hidden Z with respect to X and the corresponding representative submatrix of Y.*

Proof. This result follows from Lemma 2.1. By Lemma 2.1, we have an index set $\alpha \subseteq \{1, 2, \ldots, k\}$ and a representative submatrix V of Y such that

$$W = \begin{bmatrix} X_{\alpha\,\alpha} & X_{\alpha\,\bar{\alpha}} \\ V_{\bar{\alpha}\,\alpha} & V_{\bar{\alpha}\,\bar{\alpha}} \end{bmatrix}$$

is a K-matrix. Let G be the corresponding representative submatrix of M. Since $W \in K$, it follows that $W^t \in K$ and there is a $v \in R^k$, $v \geq 0$ such that $v^t W > 0$. Let $v = (v_\alpha, v_{\bar{\alpha}})$. Take $r(G)^t = (v_\alpha^t, 0)$ and $s(G)^t = (0, v_{\bar{\alpha}}^t)$. It is easy to verify that $GX = V$ and $r(G)^t X + s(G)^t V = v^t W > 0$. This shows that G is a hidden Z-matrix, and this completes the proof of the proposition. □

Remark 2.2. The above proposition implies in particular that if M is a vertical block hidden Z-matrix with X and Y as any matrices satisfying the conditions of Definition 2.4 then there exists a nonnegative matrix $U \in R^{k\times m}$ of the form

$$U = \begin{bmatrix} u^1 & \ldots & 0 \\ \vdots & \ldots & \vdots \\ 0 & \ldots & u^k \end{bmatrix},$$

where $u^r = (u_1^r, \ldots, u_{m_r}^r)$ is a nonnegative row vector of order m_r such that UM is a square hidden Z-matrix.

Remark 2.3. It is easy to see that if X is a K-matrix then UM is a hidden Z-matrix for any nonnegative U of the above form. For similar results on vertical block P-matrices, see [1]; for vertical block P_0- and Z-matrices, see [5].

3. Least element property. In this section, we consider the least element property of vertical block hidden Z-matrices.

DEFINITION 3.1. *Let $S \subseteq R^n$ be a polyhedral set. We say that $x \in S$ is the* least element of S with respect to the cone ordering induced by a convex cone C *if* $y - x \in C$ *for any* $y \in S$.

DEFINITION 3.2. $S \subset R^n$ *is called a* meet semisublattice (under the componentwise ordering of R^n) *if for any two vectors* $x, y \in S$ *their meet* $z = \min(x, y) \in S$.

In what follows, let M be a vertical block hidden Z-matrix with X and Y as any matrices satisfying the conditions of Definition 2.4. Note that by Lemma 2.1, X is nonsingular. Let $S = \{ v \in R^n : Xv \geq 0,\ q + Yv \geq 0 \}$.

LEMMA 3.1. *A vector* $z \in FEA(q, M)$ *iff* $v = X^{-1} z \in S$. *Also S is a meet semisublattice.*

Proof. To show this, note that $MX = Y$ and $w = q + Mz \geq 0$ as $z \in \mathrm{FEA}(q, M)$. Let $v = X^{-1}z$. So, $z = Xv \geq 0$. Note that $q + Mz = q + MXv = q + Yv \geq 0$. Hence $v \in S$.

Now given $v \in S$, take $z = Xv$. Note that $z \geq 0$. We have $q + Yv = q + MXv = q + Mz \geq 0$. Hence $z = Xv \in \mathrm{FEA}(q, M)$.

Now, we have to show that S is a meet semisublattice. Let $v^*, \bar{v} \in S$ and let $\hat{v}$ be a vector whose ith coordinate is defined by $\hat{v}_i = \min(v_i^*, \bar{v}_i)$.

Suppose $s \in J_i$, the set of indices of rows of M in the ith block. Note that

$$\begin{aligned}
q_s + (Y\hat{v})_s &= q_s + \sum_{j=1}^{k} y_{sj}\, \hat{v}_j \\
&= q_s + y_{si}\, \hat{v}_i + \sum_{j \neq i} y_{sj}\, \hat{v}_j \\
&= q_s + y_{si}\, v_i^* + \sum_{j \neq i} y_{sj}\, \hat{v}_j, \text{ assuming (without loss of generality) } \hat{v}_i = v_i^*, \\
&\geq q_s + y_{si}\, v_i^* + \sum_{j \neq i} y_{sj}\, v_j^*, \text{ since } y_{sj} \leq 0 \text{ for } j \neq i, \\
&= q_s + \sum y_{sj}\, v_j^* \geq 0, \text{ since } v^* \in S.
\end{aligned}$$

Similarly, we can show that $z = X\hat{v} \geq 0$. Thus S is a meet semisublattice. This completes the proof of Lemma 3.1. □

LEMMA 3.2. *S contains a least element.*

Proof. It is sufficient to verify that S is bounded below as S is a meet semisublattice.

Let $v \in S$ and $\tilde{q} = \begin{bmatrix} 0 \\ q_{\bar{\alpha}} \end{bmatrix}$, where $\bar{\alpha}$ is as in Lemma 2.1. Let W be as in Lemma 2.1. Note that $W^{-1} \geq 0$ and by the definition of S, we have $Xv \geq 0$ and $q + Yv \geq 0$. Hence $\tilde{q} + Wv \geq 0$. Let $u = \tilde{q} + Wv$. Then $W^{-1}u = W^{-1}\tilde{q} + v \geq 0$ as $u \geq 0$ and $W^{-1} \geq 0$. Hence $v \geq -W^{-1}\tilde{q}$. This concludes the proof. □

THEOREM 3.1. *Suppose that $M \in R^{m \times k}$ is a vertical block hidden Z-matrix of type $(m_1, m_2, \ldots, m_k)$. Then there exists a simplicial cone C in R^n such that $\forall\ q \in Pos(I, -M)$, $FEA(q, M)$ contains a least element $\bar{z}$ with respect to the cone ordering induced by C and $\bar{z}$ satisfies $\bar{z}_i \prod_{s=1}^{m_i} (q_s^i + (M^i \bar{z})_s) = 0\ \forall\ i = 1, 2, \ldots, k$.*

Proof. By Lemma 3.2, S has a least element $\bar{v}$ with respect to Pos(I). Let $\bar{z} = X\,\bar{v}$. Note that by Lemma 3.1, $\bar{z} \in \text{FEA}(q, M)$, and it follows that it is a least element of FEA(q, M) with respect to the cone ordering induced by Pos(X). Now it remains to verify that $\bar{z}_i \prod_{s=1}^{m_i}(q_s^i + (M^i\,\bar{z})_s) = 0$. To see this, we first show that if $(X\,\bar{v})_i > 0$ then $\exists$ an $s \in J_i$ such that

$$q_s + (Y\,\bar{v})_s = 0.$$

Suppose $\forall\ s \in J_i$,

$$q_s + (Y\,\bar{v})_s > 0.$$

Now consider a $v^*(\epsilon)$ whose coordinates are defined as follows:

$$v_j^*(\epsilon) = \bar{v}_j,\ \ j \neq i,$$

$$v_i^*(\epsilon) = \bar{v}_i - \epsilon.$$

Note that as X is a Z-matrix, for ϵ sufficiently small, $X\,v^*(\epsilon) \geq 0$. Also, it is easy to verify using the fact that Y is a vertical block Z-matrix that

$$q_s + (Y\,v^*(\epsilon))_s \geq 0,\ \ \forall\ s.$$

This, however, contradicts the minimality of $\bar{v}$ and completes the proof. □

We shall now prove the converse of Theorem 3.1.

THEOREM 3.2. *Suppose X is a $k \times k$ nonsingular matrix. Let $C = Pos(X)$. Suppose M is a given vertical block matrix. If FEA$(q, M) \neq \phi$ implies that FEA(q, M) has a least element with respect to the ordering induced by C, which is also a solution to the VLCP(q, M), then M is a vertical block hidden Z-matrix.*

Proof. Let $\tilde{e}^j$ be an $m \times 1$ vector whose ith coordinate $(\tilde{e}^j)_i = 1\ \ \forall\ i \in J_j$ and 0 otherwise. Also, let e_j^* be the unit vector in R^k with $(e_j^*)_j = 1$ and $(e_j^*)_i = 0$ for $i \neq j$. Now let $q^j = \tilde{e}^j - M\,e_j^*$. Clearly, $e_j^* \in \text{FEA}(q^j, M)$ and hence FEA$(q^j, M) \neq \phi$. Therefore, by our hypothesis it has a least element $\bar{z}^j$ which satisfies VLCP condition (1.4). Clearly, e_j^* does not satisfy this condition. Hence $\bar{z}^j \neq e_j^*$ and, by the minimality of $\bar{z}^j$, we have $X^{-1}(\bar{z}^j) \leq X^{-1}(e_j^*)$.

Let $v^j = X^{-1}(e_j^* - \bar{z}^j)$. Note that $0 \neq v^j \geq 0$. Now for $i \in \{1, 2, \ldots, k\} \setminus \{j\}$, we have $X_{i.}\,v^j = (e_j^* - \bar{z}^j)_i \leq 0$. Let $Y = M\,X$. Note that Y is a vertical block matrix. Now consider $Y_{s.}\,v^j$:

$$\begin{aligned} Y_{s.}\,v^j &= (Y\,v^j)_s \\ &= (M\,X\,v^j)_s \\ &= [M\,(e_j^* - \bar{z}^j)]_s \\ &= (\tilde{e}^j - q^j - M\bar{z}^j)_s \\ &= -(q^j + M\bar{z}^j)_s \text{ for } s \notin J_j. \end{aligned}$$

Therefore, noting that $(q^j + M\bar{z}^j) \geq 0$, we have $Y_{s.}\,v^j \leq 0$ for $s \notin J_j$.

Let $W = (v^1, v^2, \ldots, v^k)$. Then it follows that $\tilde{X} = X\,W$ is a Z-matrix and $\tilde{Y} = Y\,W$ is a vertical block Z-matrix.

We now have to show the existence of nonnegative vectors r and s satisfying condition (ii) of Definition 2.4. To do this consider the linear programming problem

$$\text{Minimize } e^t\, u$$

subject to

$$X\, u \;\geq\; 0,$$

$$Y\, u \;\geq\; 0,$$

where e is a k-vector of 1.

Note that u is feasible to the above problem if and only if $X\, u \in \text{FEA}(0, M)$. As $0 \in \text{FEA}(0, M)$ it follows that $\text{FEA}(0, M) \neq \phi$, and hence it has a least element under the cone ordering induced by $\text{Pos}(X)$, which is also a solution to the $\text{VLCP}(0, M)$. Therefore, the above problem has an optimal solution. By the duality theorem, there exist nonnegative vectors r and s such that $X^t r + Y^t s = e$.

As $W \geq 0$ and no column of W is 0, we have

$$\tilde{X}^t\, r \;+\; \tilde{Y}^t\, s \;=\; W^t\, (X^t\, r \;+\; Y^t\, s) \;=\; W^t\, e \;> 0.$$

This completes the proof. □

Remark 3.1. In view of Theorem 3.1, the $\text{VLCP}(q, M)$ with a vertical block hidden Z-matrix with respect to X and Y can be formulated as the linear programming problem

$$\text{Minimize} \sum_{i=1}^{k} p_i\, z_i,$$

$$w \;-\; Mz \;=\; q,$$

$$w \geq 0, \;\; z \geq 0,$$

where $p = (p_1, p_2, \ldots, p_k)$ is any vector such that $p^t\, X \;>\; 0$.

Remark 3.2. Thus the remarks of Cottle, Pang, and Stone [2, p. 212] in the context of hidden Z-matrices also apply to the vertical block hidden Z-matrices. Thus, given an arbitrary vertical block matrix M it is not in general easy to test whether or not it is vertical block hidden Z.

4. Vertical block hidden K-matrices.

DEFINITION 4.1. *Let M be a vertical block hidden Z-matrix. We say that M is a* vertical block hidden K-matrix *if every representative submatrix of M is a P-matrix.*

In the example below we exhibit the blocks by separating them from one another using blank space.

Example 4.1. Let M be the following vertical block matrix:

$$\begin{bmatrix} 1.76 & 0.36 & 0.16 \\ 1 & 0 & 0 \\ 0.80 & -0.20 & -0.20 \\ \\ 0.32 & 1.52 & 0.12 \\ 0.44 & 1.84 & 0.04 \\ \\ -1.56 & -1.16 & 0.04 \\ -0.60 & -0.60 & 0.40 \end{bmatrix},$$

where $m_1 = 3$, $m_2 = 2$, and $m_3 = 2$.

It is easy to verify that M is a vertical block hidden K-matrix with respect to X, Y,

$$\text{where } X = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -2 & 7 \end{bmatrix} \text{ and } Y = \begin{bmatrix} 3 & -1 & -1 \\ 2 & -1 & -1 \\ 2 & -1 & -2 \\ -1 & 4 & -1 \\ -1 & 5 & -2 \\ -2 & -2 & 3 \\ -1 & -2 & 4 \end{bmatrix}.$$

We take $r^t = \begin{bmatrix} 3 & 2 & 1 \end{bmatrix}$ and $s^t = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$.

The following theorem characterizes a vertical block hidden K-matrix M assuming that it is a vertical block hidden Z-matrix.

THEOREM 4.1. *Let M be a vertical block hidden Z-matrix of type $(m_1, m_2, \ldots, m_k)$. Let X and Y be as in Definition 2.4. The following are equivalent:*

(a) *M is a vertical block hidden K-matrix.*

(b) *There exists an $x \in R^k$, $x > 0$ such that $Mx > 0$.*

(c) *There exists a vector $v \in R^k$, $v > 0$ such that for any given index set $\alpha \subseteq \{1, 2, \ldots, k\}$, $Wv > 0$, where*

$$W = \begin{bmatrix} X_{\alpha\,\alpha} & X_{\alpha\,\bar{\alpha}} \\ V_{\bar{\alpha}\,\alpha} & V_{\bar{\alpha}\,\bar{\alpha}} \end{bmatrix}$$

and V is the representative submatrix of Y corresponding to any given representative submatrix G of M. Furthermore, $W \in K$.

(d) *Every representative submatrix G of M is completely hidden K; i.e., for every index set $\beta \subseteq \{1, 2, \ldots, k\}$, $G_{\beta\beta}$ is hidden K.*

Proof. (a) $\Rightarrow$ (b). Suppose M is a vertical block hidden K-matrix. In particular, by definition M is a vertical block P-matrix. Now from Theorem 6 of Cottle and Dantzig [1, p. 89] it follows that there is an $x \in R^k$, $x > 0$ such that $Mx > 0$.

(b) $\Rightarrow$ (c). Let $x > 0$, $x \in R^k$ be such that $Mx > 0$. Let $v = X^{-1}x$. We have $Xv > 0$, $Yv = MXv = Mx > 0$. By Lemma 2.1, there exists a representative submatrix V and an index set $\alpha_0 \subseteq \{1, 2, \ldots, k\}$ such that

$$W_0 = \begin{bmatrix} X_{\alpha_0\,\alpha_0} & X_{\alpha_0\,\bar{\alpha}_0} \\ V_{\bar{\alpha}_0,\,\alpha_0} & V_{\bar{\alpha}_0\,\bar{\alpha}_0} \end{bmatrix}$$

is a K-matrix. As $Xv > 0$ and $Yv > 0$, it follows that $W_0 v > 0$. This implies that $v > 0$.

Now let G be any representative submatrix of M and let H be the corresponding representative submatrix of Y. Let $\alpha \subseteq \{1, 2, \ldots, k\}$ be any index set. Consider the matrix

$$W = \begin{bmatrix} X_{\alpha\,\alpha} & X_{\alpha\,\bar{\alpha}} \\ H_{\bar{\alpha}\,\alpha} & H_{\bar{\alpha}\,\bar{\alpha}} \end{bmatrix}.$$

As $Xv > 0$, $Yv > 0$, it follows that $Wv > 0$.

Since $W \in Z$ and $v > 0$, it follows that $W \in K$.

(c) $\Rightarrow$ (d). Let G be any given representative submatrix of M. Let $\beta \subseteq \{1, 2, \ldots, k\}$ be given. By (c) the matrix

$$W = \begin{bmatrix} X_{\beta\beta} & X_{\beta\bar{\beta}} \\ V_{\bar{\beta}\beta} & V_{\bar{\beta}\bar{\beta}} \end{bmatrix}$$

is a K-matrix, where V is the representative submatrix of Y corresponding to G. We now proceed as in Theorem 3.11.19 of Cottle, Pang, and Stone [2, pp. 211–212] to conclude that every representative submatrix is completely hidden K.

(d) $\Rightarrow$ (a). Note that we have $MX = Y$ with $X \in Z$, $Y \in$ vertical block Z and $r^t X + s^t Y > 0$. Since every representative submatrix is a hidden K-matrix, it follows that every representative submatrix is a P-matrix. Hence by definition, statement (a) follows.

Remark 4.1. In relation to Remark 3.2 if we know that M is a vertical block P-matrix and wish to test its membership in vertical block hidden K then it is possible to do so by solving two linear programs: one to determine if there exists a $y > 0$ such that $My > 0$ and another to determine if the required X exists. Also the corresponding VLCP is solvable in polynomial time once we have determined the required X in polynomial time.

REFERENCES

[1] R. W. Cottle and G. B. Dantzig, *A generalization of the linear complementarity problem*, J. Combin. Theory, 8 (1970), pp. 79–90.

[2] R. W. Cottle, J. S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, New York, 1992.

[3] R. W. Cottle and A. F. Veinott, jr., *Polyhedral sets having a least element*, Math. Programming, 3 (1972), pp. 238–249.

[4] A. A. Ebiefung and M. M. Kostreva, *Z-Matrices and the Generalized Linear Complementarity Problem*, Technical Report #608, Department of Mathematical Sciences, Clemson University, Clemson, SC, 1991.

[5] A. A. Ebiefung and M. M. Kostreva, *Generalized P_0- and Z-matrices*, Linear Algebra Appl., 195 (1993), pp. 165–179.

[6] A. A. Ebiefung and M. M. Kostreva, *The generalized Leontief input-output model and its application to the choice of new technology*, Ann. Oper. Res., 44 (1993), pp. 161–172.

[7] A. A. Ebiefung and M. M. Kostreva, *Global solvability of generalized linear complementarity problems and a related class of polynomial complementarity problems*, in Recent Advances in Global Optimization, C. Floudas and P. Pardalos, eds., Princeton University Press, Princeton, NJ, 1992, pp. 102–124.

[8] T. Fujisawa and E. S. Kuh, *Piecewise-linear theory of nonlinear networks*, SIAM J. Appl. Math., 22 (1972), pp. 307–328.

[9] T. Fujisawa, E. S. Kuh, and T. Ohtsuki, *A sparse matrix method for analysis of piecewise-linear resistive networks*, IEEE Trans. Circuit Theory, 19 (1972), pp. 571–584.

[10] D. Goeleven, *A Uniqueness Theorem for the Generalized Linear Complementarity Problem*, Technical Report, Département de Mathématiques, Facultés Universitaires N.-D. de la Paix, Namur, Belgique, 1992.

[11] M. S. Gowda and R. Sznajder, *The generalized order linear complementarity problem*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 779–795.

[12] M. S. Gowda and R. Sznajder, *A generalization of the Nash equilibrium theorem on bimatrix games*, Internat. J. Game Theory, 25 (1996), pp. 1–12.

[13] G. Isac and D. Goeleven, *The implicit general order complementarity problem, models and iterative methods*, Ann. Oper. Res., 44 (1993), pp. 63–92.

[14] C. E. Lemke, *Bimatrix equilibrium points and mathematical programming*, Management Sci., 11 (1965), pp. 681–689.

[15] C. E. Lemke, *Recent results on complementarity problems*, in Nonlinear Programming, J. B. Rosen, O. L. Mangasarian, and K. Ritter, eds., Academic Press, New York, 1970, pp. 349–384.

[16] O. L. Mangasarian, *Linear complementarity problems solvable by a single linear program*, Math. Programming, 10 (1976), pp. 263–270.
[17] O. L. Mangasarian, *Generalized linear complementarity problems as linear programs*, Oper. Res. Verfahren, 31 (1979), pp. 393–402.
[18] K. G. Murty, *Linear Complementarity, Linear and Nonlinear Programming*, Sigma Series in Applied Mathematics, 3, Heldermann-Verlag, Berlin, 1988.
[19] K. P. Oh, *The formulation of the mixed lubrication problem as a generalized nonlinear complementarity problem*, Trans. ASME, J. Tribology, 108 (1986), pp. 598–604.
[20] J. S. Pang, *On cone orderings and the linear complementarity problem*, Linear Algebra Appl., 23 (1978), pp. 201–215.
[21] M. Sun, *Monotonicity of Mangasarian's iterative algorithm for generalized linear complementarity problems*, J. Math. Anal. Appl., 144 (1989), pp. 474–485.
[22] M. Sun, *Singular control problems in bounded intervals*, Stochastics, 21 (1987), pp. 303-344.
[23] B. P. Szanc, *The Generalized Complementarity Problem*, Ph.D. thesis, Rensselaer Polytechnic Institute, Troy, NY, 1989.
[24] R. Sznajder, *Degree-Theoretic Analysis of the Vertical and Horizontal Linear Complementarity Problems*, Ph.D. dissertation, Department of Mathematics, University of Maryland, Baltimore County, Baltimore, MD, 1994.
[25] R. Sznajder and M. S. Gowda, *Generalizations of P_0- and P- properties; extended vertical and horizontal linear complementarity problems*, Linear Algebra Appl., 223/224 (1995), pp. 695–715.
[26] A. TAMIR, *An application of Z-matrices to a class of resource allocation problems*, Management Sci., 23 (1976), pp. 317–323.
[27] L. Vandenberghe, B. L. De Moor, and J. Vandewalle, *The generalized linear complementarity problem applied to the complete analysis of resistive piecewise linear circuits*, IEEE Trans. Circuits and Systems, 11 (1989), pp. 1382–1391.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 191–222, January 1997

015

STABILITY OF AUGMENTED SYSTEM FACTORIZATIONS IN INTERIOR-POINT METHODS*

STEPHEN WRIGHT†

Abstract. Some implementations of interior-point algorithms obtain their search directions by solving symmetric indefinite systems of linear equations. The conditioning of the coefficient matrices in these so-called augmented systems deteriorates on later iterations, as some of the diagonal elements grow without bound. Despite this apparent difficulty, the steps produced by standard factorization procedures are often accurate enough to allow the interior-point method to converge to high accuracy. When the underlying linear program is nondegenerate, we show that convergence to arbitrarily high accuracy occurs, at a rate that closely approximates the theory. We also explain and demonstrate what happens when the linear program is degenerate, where convergence to acceptable accuracy (but not arbitrarily high accuracy) is usually obtained.

Key words. interior-point methods, symmetric indefinite matrices

AMS subject classifications. 65G05, 65F05, 90C05

PII. S0895479894271093

1. Introduction.

We focus on the core linear algebra operation in primal-dual interior-point methods for linear programming: the solution of a system of linear equations whose coefficient matrix is large, sparse, and symmetric. In existing codes, the linear system is formulated in two different ways. One formulation, usually called the *augmented system formulation*, has a symmetric indefinite coefficient matrix. The other involves a more compact (but generally denser) symmetric positive-definite matrix. A diagonal matrix D is involved in both formulations, where D has the disconcerting property that some of its elements grow to ∞ as the iterates approach the solution set. This blowup in D can produce ill conditioning in the coefficient matrix of the linear system. In this paper, we examine the augmented system and look at how various factorization algorithms for this system behave as this ill conditioning develops.

We restrict our study to three standard factorization algorithms — the Bunch–Parlett, Bunch–Kaufman, and sparse Bunch–Parlett algorithms. The last of these has been used in at least one practical interior-point code for linear programming (see Fourer and Mehrotra [5]). We assume that no attempt is made to improve the conditioning of the underlying linear systems by guessing whether each component of the solution is at a bound. Preprocessing of this kind detracts from the intuitive appeal of interior-point algorithms, namely, that they avoid explicit guessing about the contents of the basis.

In numerical experiments with feasible linear programs, we find that two distinct scenarios arise.

* Received by the editors July 13, 1994; accepted for publication (in revised form) by L. Kaufman February 5, 1996. This work was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Computational and Technology Research, U.S. Department of Energy, under contract W-31-109-Eng-38.

http://www.siam.org/journals/simax/18-1/27109.html

† Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439 (wright@mcs.anl.gov).

1. Even when the iterates are very close to the solution set, the computed search directions are good enough to produce rapid convergence of the algorithm at nearly the rates predicted by the theory. This performance is a little surprising. Since the matrix is poorly conditioned, we might have expected the computed directions to be too inaccurate to allow the algorithm to make much progress. This scenario usually occurs when the underlying linear program has a unique primal-dual solution.
2. Near the solution, calculation of the search direction fails because of breakdown of the matrix factorization, or else the computed search direction is so inaccurate that the interior-point method can move only a tiny distance along it before violating a bound. This scenario usually occurs when the underlying linear program is degenerate.

Our analysis in this paper explains these observations through a close examination of the behavior of factorization algorithms on the highly structured matrices that arise in our application. The effects of roundoff error are tracked by using fairly standard techniques from backward error analysis.

The most successful interior-point methods for practical linear programming problems are primal-dual methods. The best-known potential-reduction algorithm in this class was devised by Kojima, Mizuno, and Yoshise [9]; the review paper of Todd [17] contains a wealth of historical information on potential-reduction methods. Early developments in path-following methods are surveyed by Gonzaga [7], while Mizuno, Todd, and Ye [15] describe an important variant of these methods that does not require the iterates to stay within a cramped neighborhood of the central path. Zhang [25] extended the path-following approach further, allowing the iterates to be infeasible while retaining global convergence and polynomial complexity; see also Wright [21]. Some of these developments took place in the context of linear complementarity, a class of problems that includes linear programming as a special case.

On the computational side, the OB1 code described by Lustig, Marsten, and Shanno [10] generated search directions of the type described in this paper. They compute the maximum step α^* that could be taken along this direction without violating the positivity bounds, then set the actual step length to $.995\,\alpha^*$. Mehrotra's [14] predictor-corrector search direction differs from the one analyzed in this paper, but under our assumptions below, the difference vanishes as the solution is approached. Newer codes, such as those described by Mehrotra [14], Fourer and Mehrotra [5], Lustig, Marsten, and Shanno [12], Vanderbei [18], and Xu, Hung, and Ye [23], all implement Mehrotra's predictor-corrector strategy. These newer codes continue to use step lengths based on α^*; hence, we pay particular attention to the effect of roundoff error on this quantity.

Previous analysis of the ill-conditioned linear systems that arise in interior-point and barrier methods has been carried out by Ponceleón [16] and Wright [22]. Ponceleón [16] showed that these systems are not too sensitive to structured perturbations from a certain class provided that the underlying optimization problem is well conditioned. Wright [22] analyzed Gaussian elimination in the context of interior-point algorithms for linear complementarity problems.

Simultaneously with the original version of this paper, and independently, Forsgren, Gill, and Shinnerl [4] performed an analysis of the augmented system in barrier algorithms. Their analysis tends to be more detailed than ours, and a few of the results overlap. However, they assume that the factorization algorithms select the large diagonal elements as pivots before any others, a pattern that does not generally

occur in practice.

Vavasis [19] gives an illuminating discussion of the augmented system in contexts other than optimization. He presents a solution method that is provably stable in a certain sense, but which is not guaranteed to produce "useful" steps in the sense of this paper. Duff [3] also discusses augmented systems in a general context and describes a sparse factorization procedure.

2. Interior-point methods. We consider the linear program in standard form:

$$\min c^T x, \qquad Ax = b, \qquad x \geq 0, \tag{1}$$

where $x \in \mathsf{R}^n$ and $b \in \mathsf{R}^m$. The dual of (1) is

$$\max b^T \lambda, \qquad A^T \lambda + s = c, \qquad s \geq 0, \tag{2}$$

where $s \in \mathsf{R}^n$ and $\lambda \in \mathsf{R}^m$. A vector triple (λ^*, x^*, s^*) is a primal-dual solution if x^* is feasible for (1), (λ^*, s^*) is feasible for (2), and s^* and x^* are complementary; that is,

$$x^{*T} s^* = c^T x^* - b^T \lambda^* = 0. \tag{3}$$

We denote the set of primal-dual solutions by $\mathcal{S}$.

Each iterate (λ, x, s) of a primal-dual interior-point method satisfies the strict inequality $(x, s) > 0$. Search directions are found by applying a modification of Newton's method to the following system of nonlinear equations:

$$Ax - b = 0, \qquad A^T \lambda + s - c = 0, \qquad XSe = 0, \tag{4}$$

where $X = \operatorname{diag}(x_1, x_2, \ldots, x_n)$ and $S = \operatorname{diag}(s_1, s_2, \ldots, s_n)$. Specifically, the search direction $(\Delta\lambda, \Delta x, \Delta s)$ satisfies the linear equations

$$\begin{bmatrix} 0 & A^T & I \\ A & 0 & 0 \\ S & 0 & X \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta\lambda \\ \Delta s \end{bmatrix} = \begin{bmatrix} -A^T\lambda - s + c \\ b - Ax \\ -XSe + \sigma\mu e \end{bmatrix}, \tag{5}$$

where $\sigma \in [0, 1]$ is known as the centering parameter and the duality measure μ is defined by

$$\mu = x^T s / n.$$

The step length α along the search direction is determined by various factors; minimally, the updated x and s components are required to stay strictly positive:

$$(x, s) + \alpha(\Delta x, \Delta s) > 0. \tag{6}$$

At least half the components of (x, s) — the *critical* components — become very close to their lower bound of zero during the later stages of the algorithm. Despite this property, the step length α can be quite close to one without violating the property (6) when the search direction $(\Delta\lambda, \Delta x, \Delta s)$ is an *exact* solution of (5). If perturbations caused by roundoff are present in the critical components of $(\Delta\lambda, \Delta x, \Delta s)$, the requirement (6) can severely curtail the allowable step length and slow the convergence. Hence, it is important that the critical components of $(\Delta\lambda, \Delta x, \Delta s)$ be computed to high relative accuracy. This point provides the focus for much of our error analysis.

Throughout the paper we use $\mathbf{u}$ to denote unit roundoff, which we define implicitly by the statement that when x and y are any two floating-point numbers, op denotes $+, -, \times, /$, and $fl(z)$ denotes the floating-point approximation of any real number z, we have

$$fl(x \operatorname{op} y) = (x \operatorname{op} y)(1+\delta), \qquad |\delta| \leq \mathbf{u}. \tag{7}$$

Since our concern is with the internal workings of a single interior-point iterate, we omit iteration counters from all quantities. For this reason, we use the order notation $O(\cdot)$ in a slightly unconventional way. When ξ and η are two nonnegative numbers, we write $\xi = O(\eta)$ if there is a positive constant C (not too large) such that $\xi \leq C\eta$. We say that a matrix or vector is $O(\eta)$ if its norm is $O(\eta)$. We say that $\xi = \Omega(\eta)$ if $\xi = O(\eta)$ and $\eta = O(\xi)$.

For the purposes of this paper, we are mainly interested in how the factorizations behave relative to μ and $\mathbf{u}$. The dimensions m and n are ignored in our use of the notation $O(\cdot)$.

If G is a matrix, $G_{\cdot j}$ denotes its jth column, while $G_{i\cdot}$ denotes the ith row. The matrix whose elements are $|G_{ij}|$ is denoted by $|G|$.

We use $\|\cdot\|$ to denote any one of the equivalent matrix norms $\|\cdot\|_1$, $\|\cdot\|_2$, or $\|\cdot\|_\infty$. When G is rectangular, the 2-norm condition number is defined as follows.

DEFINITION 2.1. *Let G be a rectangular matrix with full rank, and suppose that* $\mathbf{svmax}\ (G)$ *and* $\mathbf{svmin}\ (G)$ *denote the largest and smallest singular values of G, respectively. The 2-norm condition number of G is*

$$\kappa(G) = \frac{\mathbf{svmax}\ (G)}{\mathbf{svmin}\ (G)}.$$

If G is square and nonsingular, this definition coincides with the usual definition

$$\kappa(G) = \|G\|_2 \|G^{-1}\|_2.$$

3. Definitions and assumptions. We assume throughout that the problems (1), (2) are feasible; that is, there exists at least one triple (λ, x, s) satisfying the constraints $Ax = b$, $A^T\lambda + s = c$, $(x, s) \geq 0$. Feasibility implies existence of solutions to (1), (2). The following theorem gives another consequence of feasibility.

THEOREM 3.1. *Suppose that* (1) *and* (2) *are feasible and that (λ, x, s) is any point with $(x, s) > 0$. Then there exists a solution $(\Delta\lambda, \Delta x, \Delta s)$ to* (5).

Proof. The proof follows from section 6 of Wright [21]. See, in particular, Lemma 6.2, Theorem 6.3, and the remarks in the last two paragraphs of [21]. ∎

Note that A need not have full rank for Theorem 3.1 to hold.

The set of *basic* indices $\mathcal{B} \subset \{1, 2, \ldots, n\}$ can be defined as

$$\mathcal{B} = \{i \mid s_i^* = 0 \ \text{ for all } (\lambda^*, x^*, s^*) \in \mathcal{S}\}, \tag{8}$$

while the nonbasic set $\mathcal{N}$ is

$$\mathcal{N} = \{i \mid x_i^* = 0 \ \text{ for all } (\lambda^*, x^*, s^*) \in \mathcal{S}\}. \tag{9}$$

It is well known that $\mathcal{B}$ and $\mathcal{N}$ form a partition of $\{1, 2, \ldots, n\}$ and that there is at least one solution (λ^*, x^*, s^*) that is strictly complementary, that is, $x^* + s^* > 0$ (Goldman and Tucker [6]). The cardinality of $\mathcal{B}$ is denoted by $|\mathcal{B}|$. By partitioning the columns of A according to $\mathcal{B}$ and $\mathcal{N}$, we define

$$B = [A_{\cdot j}]_{j \in \mathcal{B}}, \qquad N = [A_{\cdot j}]_{j \in \mathcal{N}}, \tag{10}$$

so that B is $m \times |\mathcal{B}|$ and N is $m \times |\mathcal{N}|$. We say that the linear program is *nondegenerate* if $|\mathcal{B}| = m$ and the primal-dual solution is unique. We assume also that B is reasonably well conditioned in nondegenerate problems.

We do not confine our analysis to one specific primal-dual algorithm. Rather, we rely on a set of assumptions that is satisfied by a variety of algorithms. The first of these assumptions concerns the iterates, the search directions, and the relationship between μ and the current infeasibility.

Assumption 1. The sequence of iterates (λ, x, s) generated by the interior-point algorithm satisfies the following properties when μ becomes sufficiently small:

$$x_i = \Omega(1) \quad (i \in \mathcal{B}), \qquad s_i = \Omega(1) \quad (i \in \mathcal{N}), \tag{11a}$$

$$x_i = \Omega(\mu) \quad (i \in \mathcal{N}), \qquad s_i = \Omega(\mu) \quad (i \in \mathcal{B}). \tag{11b}$$

In addition, the infeasibility is $O(\mu)$; that is,

$$b - Ax = O(\mu), \qquad c - A^T\lambda - s = O(\mu). \tag{12}$$

Assumption 1 is not very strong. Güler and Ye [8] study algorithms in which all iterates are strictly feasible; that is,

$$Ax = b, \qquad A^T\lambda + s = c, \qquad (x, s) > 0. \tag{13}$$

In fact, they require x and s to be slightly separated from the boundary of the positive orthant in the sense that

$$x_i s_i \geq \gamma\mu, \qquad i = 1, 2, \ldots, n \tag{14}$$

for some constant $\gamma \in (0, 1)$. They show that all limit points of such algorithms are strictly complementary solutions of (1), (2) and that most path-following and potential-reduction algorithms do in fact satisfy (14). It is easy to infer from their results that (11) holds for all subsequences that approach these limit points. Moreover, (12) is trivially satisfied for all feasible algorithms.

The infeasible-interior-point algorithm described by Wright [20] satisfies Assumption 1. So does the algorithm in [21], provided that the sequence of iterates (x, s) is bounded. Implemented algorithms such as those of Vanderbei [18], Lustig, Marsten, and Shanno [10, 11], and Xu, Hung, and Ye [23] usually step a fixed multiple of the distance to the boundary rather than enforce a potential reduction condition or a condition like (14). Nevertheless, the iteration sequence usually satisfies the properties of Assumption 1 for most practical problems.

Finally, we state without proof a technical lemma for use in later sections.

LEMMA 3.2. *Let H be a square matrix partitioned as*

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix},$$

where H_{11} and H_{22} are also square. Suppose that H_{11} and $H_{22} - H_{21}H_{11}^{-1}H_{12}$ are nonsingular. Then H is nonsingular and

$$H^{-1} = \begin{bmatrix} H_{11}^{-1} + H_{11}^{-1}H_{12}(H_{22} - H_{21}H_{11}^{-1}H_{12})^{-1}H_{21}H_{11}^{-1} & -H_{11}^{-1}H_{12}(H_{22} - H_{21}H_{11}^{-1}H_{12})^{-1} \\ -(H_{22} - H_{21}H_{11}^{-1}H_{12})^{-1}H_{21}H_{11}^{-1} & (H_{22} - H_{21}H_{11}^{-1}H_{12})^{-1} \end{bmatrix}.$$

4. Exact and approximate search directions. By defining $r_b = Ax - b$ and $r_c = A^T\lambda + s - c$ in (5), we obtain

$$\begin{bmatrix} 0 & A & 0 \\ A^T & 0 & I \\ 0 & S & X \end{bmatrix} \begin{bmatrix} \Delta\lambda \\ \Delta x \\ \Delta s \end{bmatrix} = \begin{bmatrix} -r_b \\ -r_c \\ -XSe + \sigma\mu e \end{bmatrix}. \tag{15}$$

By eliminating Δs from this system, we obtain the augmented system formulation:

$$\begin{bmatrix} 0 & A \\ A^T & -X^{-1}S \end{bmatrix} \begin{bmatrix} \Delta\lambda \\ \Delta x \end{bmatrix} = \begin{bmatrix} -r_b \\ -r_c + s - \sigma\mu X^{-1}e \end{bmatrix}, \tag{16a}$$

$$\Delta s = -s + \sigma\mu X^{-1}e - X^{-1}S\Delta x. \tag{16b}$$

In Wright [22], we performed an error analysis on a system like (16a), but in the context of a specific path-following algorithm for the monotone linear complementarity problem. Some of our results from [22] are relevant to the present case of (16), as we discuss later.

Potential difficulties with the formulation (16) arise from two sources — the possible rank deficiency in certain submatrices of A and the fact that some diagonal elements of $X^{-1}S$ and $S^{-1}X$ approach zero while others approach $+\infty$. Despite the effects of ill conditioning and finite precision, we find that the approximate search directions obtained from (16) by using standard factorization procedures are often remarkably good. They allow the interior-point algorithm to take near-unit steps and to make substantial improvements in the duality measure μ. In the following theorem, we specify a set of conditions for which this happy situation holds. In later sections, we identify situations under which these conditions hold.

In the remainder of the paper, we use α^* to denote the largest number in $[0,1]$ such that

$$(x + \alpha\Delta x, s + \alpha\Delta s) \geq 0 \quad \text{for all } \alpha \in [0, \alpha^*], \tag{17a}$$

$$(x + \alpha\Delta x)^T(s + \alpha\Delta s) \quad \text{is decreasing for } \alpha \in [0, \alpha^*]. \tag{17b}$$

THEOREM 4.1. *Suppose that Assumption* 1 *holds. Let* $(\Delta\lambda, \Delta x, \Delta s)$ *be the exact solution of* (5) (*equivalently,* (16)), *and let* $(\widehat{\Delta\lambda}, \widehat{\Delta x}, \widehat{\Delta s})$ *be an approximation to this step. Suppose that the centering parameter* σ *in* (5) *lies in the range* $[0, 1/2]$ *and that the following conditions hold:*

$$(\Delta x, \Delta s) = O(\mu), \tag{18a}$$

$$(\Delta s_N, \Delta x_B) - (\widehat{\Delta s}_N, \widehat{\Delta x}_B) = O(\mathbf{u}), \tag{18b}$$

$$(\Delta s_B, \Delta x_N) - (\widehat{\Delta s}_B, \widehat{\Delta x}_N) = O(\mu\mathbf{u}). \tag{18c}$$

Define α^* *as in* (17), *and suppose* $\hat{\alpha}^*$ *is obtained by replacing* $(\Delta x, \Delta s)$ *with* $(\widehat{\Delta x}, \widehat{\Delta s})$ *in* (17). *Then for all* μ *sufficiently small, we have*

$$1 - \alpha^* = O(\mu), \tag{19}$$

$$\hat{\alpha}^* = \alpha^* + O(\mathbf{u}) = 1 + O(\mu) + O(\mathbf{u}), \tag{20}$$

and

$$(x + \hat{\alpha}^*\widehat{\Delta x})^T(s + \hat{\alpha}^*\widehat{\Delta s})/n = \sigma O(\mu) + O(\mu(\mu + \mathbf{u})). \tag{21}$$

Proof. From (11a) and (18a), we have

$$s_N + \alpha \Delta s_N > 0, \qquad x_B + \alpha \Delta x_B > 0 \quad \text{for all } \alpha \in [0,1],$$

so these components do not restrict the value of α^*. Since $\mathbf{u}$ is much smaller than one, we use (18b) as well to deduce that

$$s_N + \alpha \widehat{\Delta s}_N \geq s_N + \alpha \Delta s_N + \alpha(\widehat{\Delta s}_N - \Delta s_N) > 0 \quad \text{for all } \alpha \in [0,1].$$

Similarly, we can show that $x_B + \alpha \widehat{\Delta x}_B > 0$ for all $\alpha \in [0,1]$.

For the decrease condition (17b) we show that the duality gap actually decreases over the entire interval $[0,1]$ for both exact and approximate search directions, so that this condition does not play a role in determining α^* or $\hat{\alpha}^*$. For the exact direction, we have from (5), (18a), and $\sigma \in [0, 1/2]$ that

$$\begin{aligned} \frac{d}{d\alpha}(x + \alpha \Delta x)^T (s + \alpha \Delta s) &= x^T \Delta s + s^T \Delta x + 2\alpha \Delta x^T \Delta s \\ &\leq -(1-\sigma) n\mu + 2\|\Delta x\| \|\Delta s\| \\ &\leq -n\mu/2 + O(\mu^2) \end{aligned}$$

for all $\alpha \in [0,1]$. Hence, for μ sufficiently small, the duality gap is decreasing over $[0,1]$. For the approximate direction $(\widehat{\Delta x}, \widehat{\Delta s})$, this bound can be modified slightly to account for the inexactness. We omit the details, which are straightforward but messy, and state the conclusion as

$$\frac{d}{d\alpha}(x + \alpha \widehat{\Delta x})^T (s + \alpha \widehat{\Delta s}) \leq -n\mu/2 + O(\mu \mathbf{u} + \mu^2).$$

Again, we find that the duality gap is decreasing over the whole interval $\alpha \in [0,1]$.

Hence, the only condition that can bound α^* and $\hat{\alpha}^*$ away from one is (17a), and then only for the $\mathcal{N}$-components of x and the $\mathcal{B}$-components of s. In fact, α^* satisfies

$$\frac{1}{\alpha^*} = \max\left(1, \max_{i \in \mathcal{B}} -\frac{\Delta s_i}{s_i}, \max_{i \in \mathcal{N}} -\frac{\Delta x_i}{x_i}\right). \tag{22}$$

From (5), we have $x_i \Delta s_i + s_i \Delta x_i = -x_i s_i + \sigma\mu$. Hence, since $x_i s_i = \Omega(\mu)$ from (11), we have

$$-\frac{\Delta s_i}{s_i} = 1 + \frac{\Delta x_i}{x_i} - \sigma \frac{\mu}{x_i s_i} < 1 + \frac{\Delta x_i}{x_i}.$$

For $i \in \mathcal{B}$ we have from (11a) and (18a) that $|\Delta x_i / x_i| = O(\mu)$ and therefore

$$\max_{i \in \mathcal{B}} -\frac{\Delta s_i}{s_i} \leq 1 + O(\mu).$$

An identical argument can be used for the other term in (22), so we have

$$\frac{1}{\alpha^*} \leq \max(1, 1 + O(\mu)) \quad \Rightarrow \quad 1 - \alpha^* = O(\mu),$$

proving (19).

For the maximum step length $\hat{\alpha}^*$ along the approximate direction $(\widehat{\Delta x}, \widehat{\Delta s})$, we have from (18c) and (11b) that

$$\frac{\widehat{\Delta s}_i}{s_i} - \frac{\Delta s_i}{s_i} = \frac{O(\mu \mathbf{u})}{\Omega(\mu)} = O(\mathbf{u}) \quad (i \in \mathcal{B}), \qquad \frac{\widehat{\Delta x}_i}{x_i} - \frac{\Delta x_i}{x_i} = O(\mathbf{u}) \quad (i \in \mathcal{N}).$$

Hence, from (22), we have

$$\frac{1}{\hat{\alpha}^*} = \max\left(1, \max_{i \in \mathcal{B}} -\frac{\widehat{\Delta s}_i}{s_i}, \max_{i \in \mathcal{N}} -\frac{\widehat{\Delta x}_i}{x_i}\right) = \frac{1}{\alpha^*} + O(\mathbf{u}). \tag{23}$$

For all sufficiently small μ, the estimates (20) follow immediately from this last expression.

Finally, for the estimate of potential decrease (21), we have from (5) that

$$\begin{aligned} &(x + \alpha \widehat{\Delta x})^T (s + \alpha \widehat{\Delta s}) \\ &= \left[x + \alpha \Delta x + \alpha(\widehat{\Delta x} - \Delta x)\right]^T \left[s + \alpha \Delta s + \alpha(\widehat{\Delta s} - \Delta s)\right] \\ &\leq n\mu(1 - \alpha(1 - \sigma)) + O(\mu \mathbf{u}) + O(\mu \mathbf{u}^2), \end{aligned} \tag{24}$$

where we have used Assumption 1 and (18) to estimate the remainder terms. Finally, we obtain (21) by substituting $\alpha = \hat{\alpha}^* = 1 + O(\mu + \mathbf{u})$ into (24). ∎

5. The augmented system. In the remainder of the paper, we focus on the procedure based on (16) for finding the search directions. In this section, we define a generalized form of the matrix in (16a) which we call a *canonical matrix*. We show that if the backward error analysis of the solution procedure satisfies a certain condition — Condition 1 below — then the approximate step $(\widehat{\Delta \lambda}, \widehat{\Delta x}, \widehat{\Delta s})$ obtained from (16) in a finite-precision environment is "useful" in the sense of Theorem 4.1.

In later sections, we define conditions under which these standard algorithms for solving symmetric indefinite systems satisfy Condition 1 and hence yield useful search directions. Our sharpened, specialized error analysis yields much stronger results than a naive application of the standard results. We also gain insight into how the algorithms work even when the nondegeneracy assumptions of sections 6, 7, and 8 fail to hold, and why they continue to generate useful search directions even for degenerate problems until μ is quite small.

Given a symmetric matrix T of order $\bar{n}$, the factorization procedures yield

$$LDL^T = PTP^T, \tag{25}$$

where P is a permutation matrix, L is unit lower triangular, and D is a block-diagonal matrix with 1×1 and 2×2 diagonal blocks. We denote the counterparts of these matrices that are actually computed in the finite-precision environment by $\hat{L}$ and $\hat{D}$, respectively.

Given the system $Tz = d$ and the data P, $\hat{L}$, and $\hat{D}$ from the factorization, we find the computed solution $\hat{z}$ by performing two vector permutations with P, triangular substitutions with $\hat{L}$ and $\hat{L}^T$, and a blockwise inversion of $\hat{D}$. Each of these operations (except the permutations) may introduce additional roundoff error, which must be accounted for in the error analysis.

For each of the methods, we focus on a single step of the factorization procedure applied to a matrix T with properties like those of our given system (16a), which we now define.

DEFINITION 5.1. *A matrix T is a* canonical matrix *if it is a symmetric permutation of*

$$\begin{bmatrix} 0 & B & N \\ B^T & 0 & 0 \\ N^T & 0 & \Lambda \end{bmatrix} + O(\mu + \mathbf{u}), \tag{26}$$

where

- *$\mu > 0$ and $\mathbf{u} \geq 0$ are small,*
- *Λ is diagonal with all diagonal elements of magnitude $\Omega(\mu^{-1})$,*
- *$B = \Omega(1)$ and $\kappa(B) = O(1)$, and*
- *$N = O(1)$.*

We call T a degenerate canonical matrix *if it has the form*

$$\begin{bmatrix} 0 & 0 \\ 0 & \Lambda \end{bmatrix} + O(\mu + \mathbf{u}), \tag{27}$$

where the zero blocks are nonvacuous.

In keeping with our particular application (16a), we use m and n to denote the number of rows and columns in the composite matrix $[B \mid N]$, respectively, and $\bar{n} = m + n$ to denote the total dimension of T.

Corresponding to our canonical matrix, we define a canonical *error* matrix. We prove that for each of the factorizations, this error matrix has the form specified in the following definition.

DEFINITION 5.2. *Let T be a canonical matrix. The corresponding* canonical error matrix *Δ is a matrix of the same dimension as T such that*

$$|\Delta| \leq \Delta_{\mathbf{u}} + |T|\delta_{\mathbf{u}}, \tag{28}$$

where $\delta_{\mathbf{u}}$ and the elements of $\Delta_{\mathbf{u}}$ are $O(\mathbf{u})$.

An important role in the pivot selection process is played by the quantities χ_i, which denote the magnitude of the largest off-diagonal element in column i, that is,

$$\chi_i = \max\{|T_{ij}| \mid j = 1, 2, \ldots, \bar{n},\ j \neq i\}. \tag{29}$$

A sufficient condition for useful steps. The following condition states the common goal of our backward error analysis of the three factorization procedures. When this condition is satisfied along with nondegeneracy of the linear program, the result of Theorem 4.1 holds.

Condition 1. Given the system $Tz = d$, where T is a canonical matrix, the symmetric factorization and solution process yields a computed solution $\hat{z}$ that satisfies

$$(T + \Delta)\hat{z} = \hat{d}, \tag{30}$$

where Δ is a canonical error matrix associated with T and $\hat{d} - d = O(\mathbf{u})$.

We allow for a perturbed right-hand side $\hat{d}$ because of the nature of our particular system (16a). The residuals r_b and r_c are computed as the difference of $O(1)$ quantities, so $O(\mathbf{u})$ perturbations will appear when they are evaluated in the obvious way. Addition of the terms s_N and $\mu X_N^{-1} e$ may give rise to errors of similar magnitude.

THEOREM 5.3. *Suppose that Assumption* 1 *holds and that the problem is nondegenerate, that is, $|\mathcal{B}| = m$, with $\kappa(B)$ moderate. Suppose that the procedure for solving*

(16) *satisfies Condition* 1, *and denote the approximate solution to* (16a) *by* $(\widehat{\Delta\lambda}, \widehat{\Delta x})$. *Then for all sufficiently small* μ, *we have*

$$(\Delta\lambda, \Delta x, \Delta s) = O(\mu) \tag{31}$$

and

$$(\Delta\lambda - \widehat{\Delta\lambda}, \Delta x_B - \widehat{\Delta x}_B) = O(\mathbf{u}), \qquad \Delta x_N - \widehat{\Delta x}_N = O(\mu\mathbf{u}). \tag{32}$$

Proof. We prove (31) by appealing to (5). By partitioning A into B and N according to (10), and partitioning the diagonal matrices S and X accordingly, we see that the matrix in (5) is a permutation of

$$\begin{bmatrix} 0 & B & N & 0 & 0 \\ B^T & 0 & 0 & I & 0 \\ N^T & 0 & 0 & 0 & I \\ 0 & S_B & 0 & X_B & 0 \\ 0 & 0 & S_N & 0 & X_N \end{bmatrix}. \tag{33}$$

Because of (11), the diagonal elements in X_B and S_N are $\Omega(1)$, while the matrices S_B and X_N are $O(\mu)$. In addition, B is square and well conditioned, so the matrix (33) is an $O(\mu)$ perturbation of a uniformly nonsingular matrix. From (5), we then have

$$(\Delta\lambda, \Delta x, \Delta s) = O(\|r_b\| + \|r_c\| + \|XSe - \sigma\mu e\|),$$

so the result (31) follows from (11) and (12).

To derive the relative error estimate (32), consider the system (16a). By permuting the matrix in accord with the $\mathcal{B} \cup \mathcal{N}$ partition, we can rewrite (16a) as follows:

$$\begin{bmatrix} 0 & B & N \\ B^T & -X_B^{-1}S_B & 0 \\ N^T & 0 & -X_N^{-1}S_N \end{bmatrix} \begin{bmatrix} \Delta\lambda \\ \Delta x_B \\ \Delta x_N \end{bmatrix} = \begin{bmatrix} -r_b \\ -(r_c)_B + s_B - \sigma\mu X_B^{-1}e \\ -(r_c)_N + s_N - \sigma\mu X_N^{-1}e \end{bmatrix}. \tag{34}$$

From (11), we have for sufficiently small μ that the diagonals in $X_B^{-1}S_B$ are $\Omega(\mu)$ while the diagonals of $X_N^{-1}S_N$ are $\Omega(\mu^{-1})$, so this coefficient matrix is canonical.

By defining

$$M_B = \begin{bmatrix} 0 & B \\ B^T & -X_B^{-1}S_B \end{bmatrix}, \qquad M_N = \begin{bmatrix} N \\ 0 \end{bmatrix},$$

$$\Lambda = -X_N^{-1}S_N, \qquad z_N = \Delta x_N, \qquad z_B = \begin{bmatrix} \Delta\lambda \\ \Delta x_B \end{bmatrix},$$

$$d_B = \begin{bmatrix} -r_b \\ -(r_c)_B + s_B - \sigma\mu X_B^{-1}e \end{bmatrix}, \qquad d_N = -(r_c)_N + s_N - \sigma\mu X_N^{-1}e,$$

we can restate the system as

$$\begin{bmatrix} M_B & M_N \\ M_N^T & \Lambda \end{bmatrix} \begin{bmatrix} z_B \\ z_N \end{bmatrix} = \begin{bmatrix} d_B \\ d_N \end{bmatrix}. \tag{35}$$

From our assumption on B, we have $M_B = O(1)$ and $M_B^{-1} = O(1)$.

Because of Condition 1, the computed solution $\hat{z}$ of (35) satisfies

$$\left(\begin{bmatrix} M_B & M_N \\ M_N^T & \Lambda \end{bmatrix} + \Delta \right) \begin{bmatrix} \hat{z}_B \\ \hat{z}_N \end{bmatrix} = \begin{bmatrix} \hat{d}_B \\ \hat{d}_N \end{bmatrix}, \tag{36}$$

where $\hat{d} - d = O(\mathbf{u})$ and the canonical error matrix Δ satisfies

$$|\Delta| \le O(\mathbf{u}) + \begin{bmatrix} |M_B| & |M_N| \\ |M_N|^T & |\Lambda| \end{bmatrix} O(\mathbf{u}) = O(\mathbf{u}) + \begin{bmatrix} 0 & 0 \\ 0 & |\Lambda| \end{bmatrix} O(\mathbf{u}).$$

By combining this estimate with (35) and (36), we obtain

$$\left(\begin{bmatrix} M_B & M_N \\ M_N^T & \Lambda \end{bmatrix} + \Delta \right) \begin{bmatrix} \hat{z}_B - z_B \\ \hat{z}_N - z_N \end{bmatrix} = -\Delta \begin{bmatrix} z_B \\ z_N \end{bmatrix} + \begin{bmatrix} \hat{d}_B - d_B \\ \hat{d}_N - d_N \end{bmatrix}. \tag{37}$$

Since $z = O(\mu)$ from (31), we have

$$\left| \Delta \begin{bmatrix} z_B \\ z_N \end{bmatrix} \right| \le \left(O(\mathbf{u}) + \begin{bmatrix} 0 & 0 \\ 0 & |\Lambda| \end{bmatrix} O(\mathbf{u}) \right) \begin{bmatrix} O(\mu) \\ O(\mu) \end{bmatrix} \le \begin{bmatrix} O(\mu\mathbf{u}) \\ O(\mathbf{u}) \end{bmatrix}, \tag{38}$$

so when we add the effect of $\hat{d} - d$, we find that the right-hand side of (37) is $O(\mathbf{u})$. For the coefficient matrix in (37) we use Lemma 3.2 with

$$\begin{aligned} H_{11} &= M_B + O(\mathbf{u}), \\ H_{12} &= M_N + O(\mathbf{u}) = O(1), \\ H_{21} &= M_N^T + O(\mathbf{u}) = O(1), \\ H_{22} &= O(\mathbf{u}) + \Lambda(I + O(\mathbf{u})) = \Lambda(I + O(\mathbf{u})). \end{aligned}$$

Lemma 3.2 yields the following estimates:

$$\begin{aligned} &(H^{-1})_{22} = \left(H_{22} - H_{21}H_{11}^{-1}H_{12}\right)^{-1} = \Lambda^{-1}(I + O(\mathbf{u} + \mu)) = O(\mu), \\ &(H^{-1})_{12} = O(\mu), \qquad (H^{-1})_{21} = O(\mu), \\ &(H^{-1})_{11} = M_B^{-1} + O(\mu + \mathbf{u}). \end{aligned}$$

By combining these observations with (38), we obtain

$$\begin{bmatrix} \hat{z}_B - z_B \\ \hat{z}_N - z_N \end{bmatrix} = \begin{bmatrix} (H^{-1})_{11} & (H^{-1})_{12} \\ (H^{-1})_{21} & (H^{-1})_{22} \end{bmatrix} O(\mathbf{u}) = \begin{bmatrix} O(\mathbf{u}) \\ O(\mu\mathbf{u}) \end{bmatrix},$$

giving (32). □

Next, we examine the accuracy of $\widehat{\Delta s}$, which is calculated by substituting $\widehat{\Delta\lambda}$ and $\widehat{\Delta x}$ into (16b).

THEOREM 5.4. *Suppose that the assumptions of Theorem* 5.3 *are satisfied and that* $\widehat{\Delta s}$ *is evaluated in floating-point arithmetic from the formula* (16b)*, with* $\widehat{\Delta x}$ *replacing* Δx*. We then have*

$$\Delta s_B - \widehat{\Delta s}_B = O(\mu\mathbf{u}), \tag{39a}$$

$$\Delta s_N - \widehat{\Delta s}_N = O(\mathbf{u}). \tag{39b}$$

Proof. Standard roundoff error analysis applied to (16b) shows that

$$\widehat{\Delta s} = -s + \sigma\mu X^{-1}e - X^{-1}S\widehat{\Delta x} + \left[|s| + \sigma\mu|X^{-1}|e + |X^{-1}S||\widehat{\Delta x}| \right] O(\mathbf{u}). \tag{40}$$

By differencing (16b) and (40), we obtain

$$|\Delta s - \widehat{\Delta s}| \le |X^{-1}S||\Delta x - \widehat{\Delta x}| + \left[|s| + \sigma\mu|X^{-1}|e + |X^{-1}S||\widehat{\Delta x}| \right] O(\mathbf{u}). \tag{41}$$

If $i \in \mathcal{B}$, we have from (11) that

$$|X_B^{-1} S_B| = O(\mu), \qquad |s_B| = O(\mu), \qquad \sigma\mu |X_B^{-1}| e = O(\mu).$$

By combining these estimates with (32) and (41), we obtain the desired result (39a). For $i \in \mathcal{N}$, we have from (11) again that

$$|X_N^{-1} S_N| = O(\mu^{-1}), \qquad |s_N| = O(1), \qquad \sigma\mu |X_N^{-1}| e = O(1),$$

while from (31) and (32) we have

$$\Delta x_N - \widehat{\Delta x}_N = O(\mu \mathbf{u}), \qquad \widehat{\Delta x}_N = O(\mu).$$

By substituting in (41), we obtain (39b). □

The last two results show that the requirements of Theorem 4.1 are satisfied, so that the algorithm can make significant progress along these search directions. We summarize the combination of Theorems 4.1, 5.3, and 5.4 as a corollary.

COROLLARY 5.5. *Suppose that Assumption* 1 *holds and that the problem is nondegenerate; that is,* $|\mathcal{B}| = m$ *with* $\kappa(B)$ *moderate. Suppose that the procedure for solving* (16) *satisfies Condition* 1. *If the approximate step is computed with* $\sigma \in [0, 1/2]$, *then for all sufficiently small* μ, *the formulae* (19), (20), *and* (21) *are satisfied.*

6. The Bunch–Kaufman factorization. We show in this section that a procedure for solving (16a) based on the Bunch–Kaufman factorization satisfies Condition 1, so that the conclusion of Corollary 5.5 applies. Since much of the analysis of this section can be reused in the analysis of the Bunch–Parlett and sparse Bunch–Parlett algorithms, we give the details here and refer to them in later sections.

It is sufficient to describe just the first stage of the procedure. Later stages apply the same technique recursively to the remaining submatrix.

The pivot selection procedure for Bunch–Kaufman [2] is as follows.

Choose $\delta \in (0, 1)$; find r such that $\chi_1 = |T_{r1}|$;
if $\chi_1 > 0$
  if $|T_{11}| \geq \delta \chi_1$
    1×1 pivot, $P_1 = I$
  else
    find χ_r;
    if $\chi_r |T_{11}| \geq \delta \chi_1^2$
      1×1 pivot, $P_1 = I$
    elseif $|T_{rr}| \geq \delta \chi_r$
      1×1 pivot; choose P_1 so that $(P_1 T P_1^T)_{11} = T_{rr}$
    else
      2×2 pivot; choose P_1 so that $(P_1 T P_1^T)_{21} = T_{r1}$
    end
  end
end.

If we denote the 1×1 or 2×2 pivot block by E and write

$$P_1 T P_1^T = \begin{bmatrix} E & C^T \\ C & \hat{T} \end{bmatrix}, \tag{42}$$

the first step of the factorization yields

$$P_1 T P_1^T = \begin{bmatrix} I & \\ CE^{-1} & I \end{bmatrix} \begin{bmatrix} E & \\ & \bar{T} \end{bmatrix} \begin{bmatrix} I & E^{-1}C^T \\ & I \end{bmatrix}, \tag{43}$$

where $\bar{T} = \hat{T} - CE^{-1}C$. The algorithm continues by applying this procedure to $\bar{T}$. Note that the χ_i are generally changed by each stage of the factorization. The submatrix CE^{-1} contains the subdiagonals in the first one or two columns of the L factor.

Bunch and Kaufman [2] show that for the particular choice $\delta = (1 + \sqrt{17})/8$, we have

$$\max_{i,j} |\bar{T}_{ij}| \leq (2.57) \max_{i,j} |T_{ij}|, \tag{44}$$

so there is a modest bound on element growth during each stage of the factorization.

When applied to canonical matrices, the Bunch–Kaufman procedure selects pivots of specific types and produces a reduced submatrix that is also canonical. We state these results in the following two theorems, whose proofs are tedious and are relegated to Appendices A.1 and A.2, respectively.

THEOREM 6.1. *Let one step of the Bunch–Kaufman factorization be applied to a canonical matrix that is not degenerate. Then*

(a) *The pivot block E will be either*
  (i) *a 1×1 block, chosen from among the diagonal elements of Λ, or*
  (ii) *a 2×2 block, in which the off-diagonal element E_{12} is one of the elements of B;*

(b) *The matrix remaining after the elimination is canonical and the absolute change in the elements of Λ is at most $O(1)$;*

(c) *Using the notation from (42), we have that $|C| = O(1)$, while*
  (i) *$|E| = O(\mu^{-1})$ and $|E^{-1}| = O(\mu)$ if E is a 1×1 pivot and*
  (ii) *$|E| = O(1)$ and $|E^{-1}| = O(1)$ if E is a 2×2 pivot.*

THEOREM 6.2. *Let one step of the Bunch–Kaufman factorization be applied to a degenerate canonical matrix. Then*

(a) *The pivot block E will be either*
  (i) *a 1×1 block, chosen from any of the diagonals (large or small), or*
  (ii) *a 2×2 block, in which all the elements are $O(\mu + \mathbf{u})$;*

(b) *The matrix remaining after the elimination is canonical (not necessarily degenerate) and the absolute change to the remaining matrix is $O(\mu + \mathbf{u})$.*

Because of Assumption 1, our initial matrix in (16a) is canonical. Barring pathological growth in the remaining submatrices, one of Theorems 6.1 and 6.2 applies at every stage of the Bunch–Kaufman factorization.

If B is *square* in the original matrix (corresponding to a nondegenerate linear program), then the remaining matrices encountered at every stage of the factorization are not degenerate. After a 1×1 pivot, the dimensions of B are unchanged, while a 2×2 pivot shrinks B by exactly one row and column, so it remains square. When a pivot causes B to disappear altogether, the reduced matrix has the form $\Lambda + O(\mu + \mathbf{u})$. It follows that in the case of square B, Theorem 6.1 is sufficient to analyze the entire factorization. The following result gives the backward error analysis for the factorization in this case.

COROLLARY 6.3. *Let the Bunch–Kaufman factorization be applied to a canonical matrix T in which B is square. Then, for all sufficiently small μ, we obtain computed factors $\hat{L}$ and $\hat{D}$ such that*

$$\hat{L}\hat{D}\hat{L}^T = PTP^T + P\bar{\Delta}P^T, \tag{45}$$

where $\bar{\Delta}$ is a canonical error matrix associated with T.

Proof. We prove the result by an induction argument on the dimension $\bar{n} = m+n$ of the matrix T. The induction is made slightly more complex than usual by the form of the canonical matrix, notably, the presence of the square matrix B of dimension $m \leq n$.

For $\bar{n} = 1$, we must have $m = 0$ and so trivially $P = 1$, $\hat{L} = 1$, $\hat{D} = T_{11}$. Therefore, (45) holds with $\bar{\Delta} = 0$.

For $\bar{n} = 2$, we have two cases $m = 0$ and $m = 1$. For $m = 0$, there are two elements of magnitude $\Omega(\mu^{-1})$ on the diagonal, while the off-diagonals are $O(\mu + \mathbf{u})$. Hence, a 1×1 pivot is chosen. If there is no pivoting, the first step of elimination yields

$$\begin{aligned}
\hat{L}_{21} &= T_{21}/T_{11} + |T_{21}/T_{11}|O(\mathbf{u}),\\
\hat{D}_{11} &= T_{11},\\
\hat{D}_{22} &= T_{22} - T_{21}^2/T_{11} + (|T_{22}| + |T_{21}^2/T_{11}|)O(\mathbf{u}).
\end{aligned}$$

Since $\hat{L}$ has unit diagonals, we obtain the following by expanding the factors:

$$\hat{L}\hat{D}\hat{L}^T = T + \begin{bmatrix} 0 & |T_{21}|O(\mathbf{u}) \\ |T_{21}|O(\mathbf{u}) & |T_{21}^2/T_{11}|O(\mathbf{u}) + |T_{22}|O(\mathbf{u}) \end{bmatrix} = T + \bar{\Delta},$$

where

$$|\bar{\Delta}| \leq |T|O(\mathbf{u}) + O(\mathbf{u}),$$

so $\bar{\Delta}$ is a canonical error matrix associated with T. The same logic applies if pivoting occurs.

In the remaining case $m = 1$, the pivot is 2×2, we have $\hat{L} = I$, $P = I$, and $\hat{D} = T$, and (45) holds trivially with $\bar{\Delta} = 0$.

We now examine a canonical matrix of dimension $\bar{n} > 2$ in which B is square and study the first stage of the factorization. Because the matrix is canonical and nondegenerate, Theorem 6.1 applies. For some permutation matrix P_1, we have from (42) and (43) that the first stage yields partial factors $\hat{L}_1$ and $\hat{D}_1$, where

$$\hat{L}_1 = \begin{bmatrix} I & 0 \\ CE^{-1} + \Delta_L & I \end{bmatrix}, \qquad \hat{D}_1 = \begin{bmatrix} E & 0 \\ 0 & \bar{T} + \Delta_D \end{bmatrix}, \tag{46}$$

and

$$|\Delta_L| \leq |C||E^{-1}|O(\mathbf{u}), \qquad |\Delta_D| \leq |\hat{T}|O(\mathbf{u}) + |C||E^{-1}||C|^T O(\mathbf{u}) = |\hat{T}|O(\mathbf{u}) + O(\mathbf{u}).$$

Note that Δ_D is a canonical error matrix corresponding to $\hat{T}$. By the proof of Theorem 6.1, the $(2,2)$ submatrix of $\hat{D}_1$ is canonical, so we use the inductive hypothesis to deduce that the $\hat{L}$, $\hat{D}$ factors of this submatrix satisfy

$$\hat{L}_2\hat{D}_2\hat{L}_2^T = P_2(\bar{T} + \Delta_D)P_2^T + P_2\bar{\Delta}_2P_2^T \tag{47}$$

for some permutation matrix P_2 and some canonical error matrix $\bar{\Delta}_2$ corresponding to $(\bar{T} + \Delta_D)$. We compose the overall factors of T as follows:

$$\hat{L} = \begin{bmatrix} I & 0 \\ P_2(CE^{-1} + \Delta_L) & \hat{L}_2 \end{bmatrix}, \qquad \hat{D} = \begin{bmatrix} E & 0 \\ 0 & \hat{D}_2 \end{bmatrix}, \qquad P = \begin{bmatrix} I & 0 \\ 0 & P_2 \end{bmatrix} P_1.$$

Now,

$$\hat{L}\hat{D}\hat{L}^T = \begin{bmatrix} E & (C+\Delta_2)^T P_2^T \\ P_2(C+\Delta_2) & \hat{L}_2\hat{D}_2\hat{L}_2^T + P_2 C E^{-1} C^T P_2^T + P_2\Delta_1 P_2^T \end{bmatrix}, \tag{48}$$

where

$$\Delta_1 = \Delta_L C^T + C\Delta_L^T + \Delta_L E \Delta_L^T, \qquad \Delta_2 = \Delta_L E,$$

and so

$$|\Delta_1| \le |C||E^{-1}||C|^T O(\mathbf{u}) = O(\mathbf{u}), \qquad |\Delta_2| \le |C||E^{-1}||E| O(\mathbf{u}) = O(\mathbf{u}).$$

By substituting (47) and (46) into (48), we obtain

$$\begin{aligned} \hat{L}\hat{D}\hat{L}^T &= \begin{bmatrix} E & (C+\Delta_2)^T P_2^T \\ P_2(C+\Delta_2) & P_2\left[\bar{T} + \Delta_D + CE^{-1}C^T + \Delta_1 + \bar{\Delta}_2\right] P_2^T \end{bmatrix} \\ &= \begin{bmatrix} E & (C+\Delta_2)^T P_2^T \\ P_2(C+\Delta_2) & P_2\left[\hat{T} + \Delta_D + \Delta_1 + \bar{\Delta}_2\right] P_2^T \end{bmatrix} \\ &= PTP^T + P\bar{\Delta}P^T, \end{aligned}$$

where

$$\bar{\Delta} = P_1^T \begin{bmatrix} 0 & \Delta_2^T \\ \Delta_2 & \Delta_D + \Delta_1 + \bar{\Delta}_2 \end{bmatrix} P_1.$$

Since $|\Delta_1| = O(\mathbf{u})$, $|\Delta_2| = O(\mathbf{u})$, and Δ_D and $\bar{\Delta}_2$ are canonical error matrices corresponding to $\hat{T}$, we have

$$\begin{aligned} |\bar{\Delta}| &\le P_1^T \begin{bmatrix} 0 & |\Delta_2|^T \\ |\Delta_2| & |\Delta_D| + |\Delta_1| + |\bar{\Delta}_2| \end{bmatrix} P_1 \\ &\le O(\mathbf{u}) + P_1^T \begin{bmatrix} 0 & 0 \\ 0 & |\hat{T}| \end{bmatrix} P_1 O(\mathbf{u}) \\ &\le O(\mathbf{u}) + |T| O(\mathbf{u}). \end{aligned}$$

Hence, $\bar{\Delta}$ is a canonical error matrix corresponding to T.

We complete the proof by noting that Theorem 6.1 can be applied to the remaining matrix because it is also canonical and nondegenerate. □

Given the system $Tz = d$ and the data P, $\hat{L}$, and $\hat{D}$ from the factorization, the computed solution $\hat{z}$ is found by performing two vector permutations with P, triangular substitutions with $\hat{L}$ and $\hat{L}^T$, and a blockwise inversion of $\hat{D}$. The 2×2 diagonal blocks in $\hat{D}$ can be handled by the Gaussian elimination procedure outlined in the following technical lemma, which is proved in Appendix A.3. It is easy to show that the elements of the pivot block E satisfy the condition (49).

LEMMA 6.4. *Consider the 2×2 linear system $Ey = g$ in which E is symmetric with*

$$|E_{11}| \le \delta |E_{12}|, \qquad |E_{11}||E_{22}| \le \delta^2 |E_{12}|^2 \tag{49}$$

for some $\delta \in (0,1)$. Then if we compute the solution by applying Gaussian elimination to the permuted system

$$\begin{bmatrix} E_{12} & E_{22} \\ E_{11} & E_{12} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} g_2 \\ g_1 \end{bmatrix}, \tag{50}$$

the computed solution $\hat{y}$ satisfies

$$(E + \Delta_E)\hat{y} = g,$$

where

$$|\Delta_E| \leq |E|O(\mathbf{u}). \tag{51}$$

The additional error that is introduced during recovery of the solution with the computed factors $\hat{L}$, $\hat{D}$, and $\hat{L}^T$ is quantified in the next result.

LEMMA 6.5. *Suppose the assumptions and notation of Corollary* 6.3 *hold. Then the computed solution $\hat{z}$ to the system $\hat{L}\hat{D}\hat{L}^T z = Pd$ satisfies*

$$(\hat{L}\hat{D}\hat{L}^T + P\hat{\Delta}P^T)\hat{z} = Pd, \tag{52}$$

where $\hat{\Delta}$ is a canonical error matrix associated with T.

Proof. From standard results for triangular substitution, the computed solution of $\hat{L}z_a = Pd$ satisfies

$$(\hat{L} + \hat{\Delta}_{L1})\hat{z}_a = Pd, \qquad |\hat{\Delta}_{L1}| \leq |\hat{L}|O(\mathbf{u}).$$

A similar result holds for triangular substitution with the transpose $\hat{L}^T$.

For the solution of $\hat{D}z_b = \hat{z}_a$, we note that $\hat{D}$ is block-diagonal with 1×1 and 2×2 blocks. For the 2×2 pivot blocks that arise in the Bunch–Kaufman procedure, the assumptions of Lemma 6.4 hold, so the computed solution $\hat{y}$ of a 2×2 subsystem $Ey = g$ satisfies

$$(E + \Delta_E)\hat{y} = g, \qquad |\Delta_E| = |E|O(\mathbf{u}). \tag{53}$$

When E is a 1×1 block, the estimate (53) holds trivially. Hence, the computed solution $\hat{z}_b$ of $\hat{D}z_b = \hat{z}_a$ satisfies

$$(\hat{D} + \hat{\Delta}_D)\hat{z}_b = \hat{z}_a, \qquad |\hat{\Delta}_D| \leq |\hat{D}|O(\mathbf{u}).$$

By combining the error expressions for the three component systems, we find that our computed solution $\hat{z}$ satisfies

$$(\hat{L} + \hat{\Delta}_{L1})(\hat{D} + \hat{\Delta}_D)(\hat{L} + \hat{\Delta}_{L2})^T\hat{z} = Pd.$$

Multiplying the matrix products, we find that (52) is satisfied with

$$P|\hat{\Delta}|P^T \leq |\hat{L}||\hat{D}||\hat{L}|^T O(\mathbf{u}) + O(\mu\mathbf{u} + \mathbf{u}^2).$$

From our earlier discussions on the composition of $\hat{L}$ and $\hat{D}$, it is easy to see that the absolute matrix product $|\hat{L}||\hat{D}||\hat{L}|^T$ contains all $O(1)$ elements, except for the large diagonals, which occur in the same positions as in PTP^T. Hence $P\hat{\Delta}P^T$ is a canonical error matrix corresponding to PTP^T, and our proof is complete. ∎

We can now summarize the effects of roundoff error on the entire solution process for (16) in the following theorem.

THEOREM 6.6. *Suppose T is a canonical matrix in which B is square. Then, for all sufficiently small μ, the Bunch–Kaufman factorization followed by the solution process outlined above satisfies Condition* 1.

Proof. As we noted immediately following Condition 1, the actual right-hand side may differ by terms of $O(\mathbf{u})$ from its "theoretical" value d. From (52), the computed solution $\hat{z}$ to $Tz = d$ satisfies

$$(\hat{L}\hat{D}\hat{L}^T + P\hat{\Delta}P^T)\hat{z} = \hat{d}.$$

Substituting from (45), we obtain

$$(PTP^T + P\bar{\Delta}P^T + P\hat{\Delta}P^T)\hat{z} = P\hat{d},$$

so Condition 1 follows when we set $\Delta = \bar{\Delta} + \hat{\Delta}$. □

We have shown that in the case of a nondegenerate linear program, the procedure based on applying Bunch–Kaufman to (16a) leads to approximate steps $(\widehat{\Delta\lambda}, \widehat{\Delta x}, \widehat{\Delta s})$ that satisfy the conditions of Theorem 4.1. The estimate (20) implies that during the final iterations of a primal-dual algorithm, near-unit steps can be taken along these directions without leaving the nonnegative orthant. Moreover, if the centering parameter σ is small or zero, a large reduction in the duality gap μ can be expected. In the extreme case $\sigma = 0$ (the "affine-scaling" choice), linear convergence with a rate constant of $O(\mathbf{u})$ can be attained if the actual step length is close to $\hat{\alpha}^*$. Most practical algorithms choose the step length to be a fixed multiple — typically .95 or .9995 — of $\hat{\alpha}^*$, and indeed these methods often converge rapidly during their final stages. For algorithms that use a more theoretically justifiable definition of step length the story is not, unfortunately, this simple. In [22, section 4], for instance, extra restrictions are applied to α to ensure that (12) and (14) continue to hold at the next iterate. These restrictions may result in α being much smaller than one. This case is analyzed in [22, section 4], so we do not repeat it here.

Finally, we note that the lower triangle L produced by the Bunch–Kaufman factorization may contain elements that are much larger than those of the original matrix T. This phenomenon has been scrutinized in a recent report by Ashcraft, Grimes, and Lewis [1], who observe that it leads to convergence difficulties in a nonlinear programming code. In the context of our canonical matrix of Theorem 6.1, this blowup problem does not occur. As we show in part (c) of the theorem, the contribution CE^{-1} made by one step of Bunch–Kaufman is either $O(\mu)$ or $O(1)$. The blowup problem may occur, however, when we have a *degenerate* canonical matrix as in Theorem 6.2. We only have to deal with matrices like this when the linear program itself is degenerate, and in this case there are other, more serious difficulties to face, as we discuss in section 9.

7. The Bunch–Parlett factorization. The Bunch–Parlett searches the entire remaining matrix for each pivot, not just one or two columns. The pivot selection procedure is as follows.

Choose $\delta \in (0,1)$, $\chi_{\text{off}} = |T_{rs}| = \max_{i \neq j} |T_{ij}|$, $\chi_{\text{diag}} = |T_{pp}| = \max_i |T_{ii}|$;
if $\chi_{\text{diag}} \geq \delta \chi_{\text{off}}$
    $s = 1$ and choose P_1 so that $(P_1 T P_1^T)_{11} = T_{pp}$
else
    $s = 2$ and choose P_1 so that $(P_1 T P_1^T)_{21} = T_{rs}$
end.

The elimination step is identical to Bunch–Kaufman, and the process of using the LDL^T factorization to solve the system $Tz = d$ is the same as in the preceding section. As in Bunch–Kaufman, the value $\delta = (1+\sqrt{17})/8$ leads to the modest bound of 2.57 on element growth at each stage.

When applied to canonical matrices, the Bunch–Parlett factorization proceeds in three stages.

1. All the diagonal elements of Λ are selected as 1×1 pivots.
2. 2×2 pivots of the type described in Theorem 6.1(a) are chosen.
3. When no more 2×2 pivots like this are available and the remaining matrix contains only elements of size $O(\mu + \mathbf{u})$, a combination of small 1×1 and 2×2 pivots is used to complete the factorization process.

We prove this assertion in the following lemma.

THEOREM 7.1. *Suppose that the Bunch–Parlett procedure is applied to a canonical matrix. Then the factorization proceeds according to the three-stage outline above. If the canonical matrix has B square and is nonvacuous, the factorization is completed by stages* 1 *and* 2; *stage* 3 *is vacuous.*

Proof. Assuming that Λ is not vacuous, we have at the pivot selection step that $\chi_{\text{off}} = O(1)$ and $\chi_{\text{diag}} = \Omega(\mu^{-1})$. The pivot element will therefore be one of the large diagonals corresponding to Λ. The remaining matrix is updated by subtracting $CE^{-1}C$, where clearly $C = O(1)$ and $E^{-1} = O(\mu)$. Hence, the remaining matrix retains canonical form.

We can apply this argument inductively until all the diagonals in Λ are exhausted. At the end of stage 1, the remaining matrix has the form

$$\begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix} + O(\mu + \mathbf{u}). \tag{54}$$

Stage 2 now begins. If B is not vacuous, we have $\chi_{\text{off}} = O(1)$ and $\chi_{\text{diag}} = O(\mu + \mathbf{u})$. In fact, by the assumption $B = \Omega(1)$, we have $\chi_{\text{off}} = \Omega(1)$, and the element T_{rs} that achieves the maximum comes from B. The 2×2 block with off-diagonal element T_{rs} is selected as the pivot. After the elimination step, the size of B is reduced by one row and column. The proof of Theorem 6.1(b) can be applied again here to show that the remaining matrix is also canonical, so 2×2 pivots of this type will continue to be selected until B vanishes.

The number of steps in stage 2 is $\min(\text{rows}(B), \text{columns}(B))$. At the end of this stage, the remaining matrix is square with dimension $|\text{rows}(B) - \text{columns}(B)|$ and all its elements have size $O(\mu + \mathbf{u})$. In stage 3, both 1×1 and 2×2 pivots may be used to factor this matrix. If B is square, the factorization is complete after stage 2. □

The other major results of section 6 continue to hold when the Bunch–Parlett algorithm is used instead of Bunch–Kaufman; only trivial adjustments to the analysis in section 6 and Appendix A.1 are necessary. We summarize the conclusions in the following theorem.

THEOREM 7.2. *Suppose T is a canonical matrix in which B is square. Then, for all sufficiently small μ, the Bunch–Parlett factorization followed by the solution process outlined in section* 6 *satisfies Condition* 1.

8. The sparse Bunch–Parlett factorization. Several authors (notably Fourer and Mehrotra [5]) have proposed a sparse variant of the Bunch–Parlett factorization that compromises between maintaining sparsity and limiting element growth in the remaining matrix. We outline the pivot selection procedure as described by [5], with a slight modification noted, below.

For each index $i = 1, 2, \ldots, \bar{n}$ we define the *degree* n_i to be the number of off-diagonal nonzeros in row i. We also define an estimate of the joint nonzero content

of rows i and j by

$$\hat{n}_{ij} = \min(n_i + n_j - 4, \bar{n} - 2).$$

A 2×2 pivot block

$$E = \begin{bmatrix} T_{ii} & T_{ij} \\ T_{ij} & T_{jj} \end{bmatrix} \tag{55}$$

is termed *oxo* if both of T_{ii} and T_{jj} are zero, *tile* if one of T_{ii} and T_{jj} is zero, and *full* if both of T_{ii} and T_{jj} are nonzero. We define a *cost* associated with using (55) as the pivot block in each of these three cases by

$$\begin{array}{ll} \textit{oxo:} & (n_i - 1)(n_j - 1), \\ \textit{tile:} & (n_i - 1)(\hat{n}_{ij} + 1) \quad \text{if } T_{ii} = 0, \qquad (n_j - 1)(\hat{n}_{ij} + 1) \quad \text{if } T_{jj} = 0, \\ \textit{full:} & \hat{n}_{ij}^2. \end{array}$$

The cost is an estimate of the fill-in associated with using (55) as the pivot block.

For prospective pivots, we define stability criteria in terms of the usual constant $\delta \in (0, 1)$ and the off-diagonal norms χ_i defined in (29). Any 1×1 pivot must satisfy

$$|T_{ii}^{-1}|\chi_i \leq 2/\delta, \tag{56}$$

while a 2×2 pivot (55) must have

$$\left| \begin{bmatrix} T_{ii} & T_{ij} \\ T_{ij} & T_{jj} \end{bmatrix}^{-1} \right| \begin{bmatrix} \chi_i \\ \chi_j \end{bmatrix} \leq \begin{bmatrix} 1/\delta \\ 1/\delta \end{bmatrix}. \tag{57}$$

The pivot selection procedure is as follows.

for $r = 1, 2, \ldots$
  for i with $n_i = r$
    consider T_{ii} with degree r;
    if any of these elements satisfy (56)
      accept as a 1×1 pivot and exit;
    else label it as unstable;
  end

  for unstable pivots T_{ii} from the previous loop
    consider 2×2 pivots involving T_{ii}, with costs at most
      $(r-1)^2$, $(r-1)(2r-3)$, and $(2r-4)^2$
      for oxo, tile, and full pivots, respectively;
    if any of these blocks satisfy (57)
      accept as a 2×2 pivot and exit;
  end
end.

The pivot selection pattern for the sparse Bunch–Parlett algorithm is essentially the same as for the Bunch–Kaufman algorithm, as described in Theorems 6.1 and 6.2. We prove this result in Appendix A.4 since the analysis differs a little from the Bunch–Kaufman case.

THEOREM 8.1. *The results of Theorems* 6.1 *and* 6.2 *hold when the sparse Bunch–Parlett factorization is used in place of the Bunch–Kaufman procedure.*

To obtain this result, we modified the acceptance condition (56) for 1×1 pivots. In the description of [5], the right-hand side is $1/\delta$ rather than $2/\delta$. With the original

choice, the sparse Bunch–Parlett algorithm applied to a degenerate canonical matrix could allow another type of pivot: a 2×2 pivot in which one diagonal is from Λ and the other has size $O(\mu + \mathbf{u})$. A pivot of this type is poorly conditioned and will generally lead to instability during the blockwise inversion of $\hat{D}$.

The other major results of section 6 also continue to hold when the sparse Bunch–Parlett algorithm is used in place of Bunch–Kaufman. We summarize the conclusions in the following theorem.

THEOREM 8.2. *Suppose T is a canonical matrix in which B is square. Then, for all sufficiently small μ, the sparse Bunch–Parlett factorization followed by the solution process outlined in section* 6 *satisfies Condition* 1.

9. The degenerate case. When the linear program (1), (2) is degenerate — $|\mathcal{B}| \neq m$ — the three factorization procedures can no longer run to completion with just the two kinds of pivots described in Theorem 6.1. The nonsquare shape of B in the matrix (34) means that pivots of size $O(\mu + \mathbf{u})$ — either 1×1 or 2×2 — are used at some point in the factorization process. The factorizations fail only if these pivots are exactly zero, which happens often on small problems but not otherwise. The more common outcome is that the interior-point algorithm makes only slow or erratic progress after μ has achieved a certain (small) value. In this section we sketch the reasons for this outcome.

In all the factorizations above, the large diagonal elements in $X_N^{-1} S_N$ are used as 1×1 pivots. Even though these pivots are not necessarily used before any others (except in the Bunch–Parlett algorithm), the factorizations behave as if they were solving the system (16) in the equivalent, partitioned form

$$\begin{aligned} &\begin{bmatrix} N X_N S_N^{-1} N^T & B \\ B^T & -X_B^{-1} S_B \end{bmatrix} \begin{bmatrix} \Delta\lambda \\ \Delta x_B \end{bmatrix} \\ &\quad = \begin{bmatrix} -r_b + N S_N^{-1} X_N [-(r_c)_N + s_N - \sigma\mu X_N^{-1} e] \\ -(r_c)_B + s_B - \sigma\mu X_B^{-1} e \end{bmatrix}, \end{aligned} \tag{58}$$

$$\Delta x_N = X_N^{-1} S_N \left[(r_c)_N - s_N + \sigma\mu X_N^{-1} e + N^T \Delta\lambda \right]. \tag{59}$$

The coefficient matrix in (58) is an $O(\mu)$ perturbation of the matrix

$$\begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}. \tag{60}$$

Since B is well conditioned by Definition 5.1, the matrix in (60) has $2\min(|\mathcal{B}|, m)$ nonzero singular values of magnitude $\Omega(1)$. In the nondegenerate case, (60) is well conditioned. Otherwise, it has $|m - |\mathcal{B}||$ zero singular values. When $|\mathcal{B}| < m$, the null space of (60) is spanned by

$$\begin{bmatrix} \bar{Z} \\ 0 \end{bmatrix}, \tag{61}$$

where $\bar{Z}$ is an $m \times (m - |\mathcal{B}|)$ matrix of full rank such that $B^T \bar{Z} = 0$. When $|\mathcal{B}| > m$, the null space of (60) is spanned by the matrix

$$\begin{bmatrix} 0 \\ \hat{Z} \end{bmatrix}, \tag{62}$$

where $\hat{Z}$ spans the null space of B. For small μ, these null spaces are not altered much by the perturbation of size $O(\mu)$ that is present in the matrix (58) because the nonzero singular values of (60) are well separated from zero. Perturbations in the solution of (58) due to roundoff will occur mainly in the space of small singular values. Hence, when $|\mathcal{B}| < m$, the perturbations occur mostly in the range space of the matrix (61), that is, in the components of $\Delta\lambda$. Similarly, when $|\mathcal{B}| > m$, the perturbations occur in the range space of the matrix (61), that is, in the components of Δx_B.

The main source of difficulty is inaccuracy in the computed residual vectors r_b and r_c which, as mentioned above, contain errors of $O(\mathbf{u})$. In the case $|\mathcal{B}| > m$, these perturbations are magnified by the inverses of the small singular values, usually leading to errors of size about $O(\mathbf{u}/\mu)$ in the components of Δx_B. The large relative errors in Δx_B induce large relative errors in Δs_B through the formula (16b). The step length to the boundary α^* may therefore be sharply curtailed because of the nonnegativity requirements (17a). In the case $|\mathcal{B}| < m$, the large relative errors in $\Delta\lambda$ induce errors in Δx_N through the formula (59), while these errors in turn induce large relative errors in Δs_N through (16b). The step length may again be curtailed as a result.

Errors from sources other than the vector r are less significant.

If we have a strictly feasible starting point (see (13)), then we can simply set $r = 0$ throughout the algorithm. In this case, we can fix r at zero in the computations and avoid the problem above. It is usually not easy to find such a starting point, however, so some thought should be given to other ways of dealing with the problem.

One option is to simply terminate the algorithm when it stalls, declaring success if both μ and r are small. This option works well for most purposes since stalling usually occurs only after μ is reduced to $O(\mathbf{u})$, by which time the problem has usually converged to acceptable accuracy. Fourer and Mehrotra [5] report that the convergence criteria are usually satisfied before the ill effects of roundoff are seen. Our testing in section 10 allows a similar conclusion.

A second option is to switch to a termination procedure when the interior-point algorithm stalls. A finite termination procedure (see, for example, Ye [24]) or crossover to the simplex method (Megiddo [13]) could be activated.

A third option is simply to fix r at zero in the computations once it has reached the $O(\mathbf{u})$ level, because at this stage our current point is feasible to within the limits of floating-point arithmetic. By doing so, we are effectively introducing a perturbation into the problem to freeze the infeasibility at its current level. This perturbation has an interesting effect: it moves the solution to a particular vertex of the previously optimal face, changing the $\mathcal{B} \cup \mathcal{N}$ partition appropriately. If we continue to run the interior-point algorithm to higher accuracy, it eventually converges to this vertex, but only after going through many more iterates (and taking some sharp turns in the process). The result of this process is similar to what we would achieve with a crossover to simplex, but the computational cost would generally be much higher.

10. Computational experiments. We report here on some computational experiments that demonstrate the effects described above. Our testbed algorithm is the infeasible-interior-point path-following algorithm described in Wright [21]. In exact arithmetic, this algorithm achieves superlinear convergence because it eventually always takes affine-scaling steps ($\sigma = 0$ in (5)) with step length α approaching one. This algorithm performs well on practical problems, but is not as fast as codes that use the Mehrotra predictor-corrector heuristic, for which no solid convergence theory exists, except in the nondegenerate case. The asymptotic behavior in finite precision

is quite similar for the two algorithms.

To show that the finite-precision effects are not confined to "nice" problems, we generate problems with fairly wide variations in the components of A, x_B^*, and s_N^*. The matrix A is dense and random, with elements defined by

$$A_{1j} = \tau 10^{6\tau-3}, \qquad j = 1, \ldots, n,$$
$$A_{ij} = (\tau - .5)10^{6\tau-3}, \qquad i = 2, \ldots, m, \quad j = 1, \ldots, n,$$

where every instance of τ is selected from a uniform distribution on the interval $[0, 1]$. (We choose all the elements in the first row of A to be positive to ensure that the feasible region is bounded.) We control the size of the index sets $\mathcal{B}$ and $\mathcal{N}$ (to control the amount of degeneracy) and set

$$\mathcal{N} = \{1, 2, \ldots, |\mathcal{N}|\}, \qquad \mathcal{B} = \{1, 2, \ldots, n\} \backslash \mathcal{N}.$$

We choose a particular solution (λ^*, x^*, s^*) by setting

$$\lambda^* = e, \qquad s_B^* = 0, \qquad x_N^* = 0,$$
$$s_i^* = 10^{4\tau-2}, \ \ i \in \mathcal{N}, \qquad x_i^* = 10^{3\tau-1}, \ \ i \in \mathcal{B},$$

where each τ is as before. The vectors b and c are determined by the choices of A and (λ^*, x^*, s^*).

The LAPACK Bunch–Kaufman factorization routines **dsytrf** and **dsytrs** are used to solve (16a). These routines (and the rest of our code) use double-precision arithmetic, giving $\mathbf{u} \approx 10^{-14}$ on the SPARC-5 on which these results were obtained.

We report on problems with $m = 6$, $n = 12$. (In problems smaller than this, exactly zero pivots often occur in degenerate cases, leading to breakdown.) Termination occurs when $\mu \leq 10^{-30}$ — an artificially stringent criterion, chosen to give us a clear look at asymptotic effects.

The first result is for a nondegenerate problem, for which $|\mathcal{B}| = m = 6$. Table 1 shows the sizes of μ and $\|r\|$ on each iterate. For the reasons that we outlined immediately following Condition 1, $\|r\|$ stabilizes at a magnitude of $O(\mathbf{u})$. The duality gap μ does not converge subquadratically (as it would in exact arithmetic) but rather exhibits extremely fast linear convergence, with a rate constant of about 10^{-10}. This is exactly the effect predicted by formula (21) for the affine-scaling steps that are taken on the last four iterations.

To see that the pivots have the properties predicted by Theorem 6.1, we examine the matrix D from the Bunch–Kaufman factorization. Table 2 shows D at iteration 17, when $\mu \approx 10^{-7}$. As expected, there are six 1×1 pivots of magnitude $\Omega(\mu^{-1})$ and six 2×2 pivots in which the diagonals are tiny and the off-diagonals are $\Omega(1)$. The same structure is present in D at every iteration after iteration 15.

Our second example is for a dual degenerate problem with $|\mathcal{B}| = 8 > m$. As can be seen from Table 3, the algorithm achieves fairly high accuracy after about 20 iterations, but no further improvement can be made after that point. The behavior is consistent with the discussion of section 9. It suggests that the results of section 6 are "tight," in that we cannot prove that "useful" search directions are obtained for arbitrarily small μ.

Examination of the D factor for the second example (Table 4) shows that the pivot pattern is in line with the predictions of Theorems 6.1 and 6.2. Together, these results imply that there are exactly $\min(m, |\mathcal{B}|)$ of the stable 2×2 pivots with an

TABLE 1
Nondegenerate problem: $m = 6, n = 12$.

k	$\log_{10}\mu_k$	$\log_{10}\|r^k\|_1$	Affine step?
1	5.4	3.1	
2	4.7	2.3	
3	4.3	1.6	
4	3.8	0.8	
5	3.1	−12.0	
⋮	⋮	⋮	
15	−3.2	−14.0	
16	−4.6	−13.7	∗
17	−7.2	−14.4	∗
18	−12.3	−14.1	∗
19	−22.1	−13.8	∗
20	−33.3	−14.2	termination

TABLE 2
The D factor at iteration 17 *of the nondegenerate test problem* ($*$ = *magnitude less than* 10^{-6}).

Row/Column	Pivot block	
1,2	∗	.94(1)
	.94(1)	∗
3,4	∗	−.91(2)
	−.91(2)	∗
5	.26(7)	
6	.30(11)	
7	.33(10)	
8	.47(7)	
9,10	−.30(−5)	.71(2)
	.71(2)	∗
11,12	−.27(−3)	−.15(2)
	−.15(2)	∗
13,14	∗	−.31(0)
	−.31(0)	−.49(−5)
15,16	∗	.16(0)
	.16(0)	∗
17	.27(4)	
18	.32(6)	

TABLE 3
Dual degenerate problem: $m = 6, n = 12, |\mathcal{B}| = 8$.

k	$\log_{10}\mu_k$	$\log_{10}\|r^k\|_1$	Affine step?
1	5.4	3.1	
⋮	⋮	⋮	
19	−6.0	−13.8	
20	−9.8	−14.1	∗
21	−13.6	−14.2	∗
22	−14.8	−13.8	∗
23	−15.4	−13.2	∗
⋮	⋮	⋮	
99	−17.5	−13.5	
100	−17.5	−13.4	
⋮	⋮	⋮	

TABLE 4
The D factor at iteration 17 of the degenerate test problem with $m = 6, n = 12, |\mathcal{B}| = 8$ ($*$ = magnitude less than 10^{-6}).

Row/Column	Pivot block	
1,2	$*$	.95(1)
	.95(1)	$*$
3,4	$*$	$-.92(2)$
	$-.92(2)$	$*$
5,6	$*$	.26(2)
	.26(2)	$*$
7	.86(23)	
8	.85(18)	
9	.55(20)	
10	.29(17)	
11,12	$*$	.71(2)
	.71(2)	$*$
13,14	$*$	$-.30(0)$
	$-.30(0)$	$*$
15,16	$*$	.15(0)
	.15(0)	$*$
17	$.20(-13)$	
18	$-.60(-19)$	

off-diagonal from B and $|\mathcal{N}| = n - |\mathcal{B}|$ of the large 1×1 pivots. Together, these stable pivots account for

$$2\min(m, |\mathcal{B}|) + |\mathcal{N}| = n + m - |m - |\mathcal{B}|| \tag{63}$$

stages of the factorization, so unstable pivots are used on the remaining submatrix whose dimension is $|m - |\mathcal{B}||$. In Table 4, we see that the last two 1×1 pivots are unstable, as expected. As we described in the first part of section 9, the errors in $\widehat{\Delta x}_B$ and $\widehat{\Delta s}_B$ are preventing further progress. On iteration 100, the computed affine step has $\|\widehat{\Delta x}_B\|_\infty = .17(6)$, while its exact counterpart would have $\|\Delta x_B\|_\infty = O(\mu)$. By comparing components of $\widehat{\Delta s}_B$ with s_B, we find that the step to the boundary is sharply curtailed by the restriction $s_B + \alpha\widehat{\Delta s}_B \geq 0$ (cf. (23)). The remaining components of the step do not contain deleterious errors; we have

$$\|\widehat{\Delta x}_N\|_\infty = .59(-18), \qquad \|\widehat{\Delta\lambda}\|_\infty = .66(-14), \qquad \|\widehat{\Delta s}_N\|_\infty = .11(-12).$$

Finally, we consider a primal degenerate problem with $|\mathcal{B}| = 4 < m$. The iteration schedule in Table 5 shows similar behavior to the dual degenerate problem. The D factor from iteration 100 is shown in Table 6. All pivots are stable except for the last two 1×1 blocks, which again matches the prediction (63). As discussed in section 9, the deleterious errors occur in the subvector $\widehat{\Delta\lambda}$, so errors are induced in $\widehat{\Delta s}_N$ and $\widehat{\Delta x}_N$ through formulas (59) and (16b). On iteration 100, we have $\|\widehat{\Delta\lambda}\|_\infty = .32(5)$ and $\|\widehat{\Delta s}_N\|_\infty = .30(7)$ for the affine-scaling step. The components $\widehat{\Delta x}_B$ and $\widehat{\Delta s}_B$ are not affected; their ∞-norms are $.17(-18)$ and $.51(-12)$, respectively.

Appendix A. Proofs of theorems from sections 6 and 8.

A.1. Proof of Theorem 6.1. We prove (a) of Theorem 6.1 by systematically excluding the other possible choices for pivots.

(iii) The pivot is 1×1 and is a diagonal element from either the $(1,1)$ or $(2,2)$ blocks of the canonical matrix. Inspection of the Bunch–Kaufman algorithm

TABLE 5
Primal degenerate problem: $m = 6, n = 12, |\mathcal{B}| = 4$.

k	$\log_{10}\mu_k$	$\log_{10}\|r^k\|_1$	Affine step?
1	5.4	3.1	
⋮	⋮	⋮	
15	−5.3	−13.9	
16	−8.8	−13.7	∗
17	−13.7	−14.2	∗
18	−14.0	−11.6	∗
⋮	⋮	⋮	
99	−17.6	−13.9	
100	−17.6	−14.0	
⋮	⋮	⋮	

TABLE 6
The D factor at iteration 17 *of the degenerate test problem with* $m = 6, n = 12, |\mathcal{B}| = 4$ ($*$ =*magnitude less than* 10^{-6}).

Row/Column	Pivot block	
1,2	∗	.95(1)
	.95(1)	∗
3	.49(23)	
4	.53(19)	
5	.58(19)	
6	.27(20)	
7	.53(9)	
8	.12(21)	
9,10	∗	.71(2)
	.71(2)	∗
11,12	∗	−.15(2)
	−.15(2)	∗
13	.25(18)	
14	.76(17)	
15,16	∗	−.16(1)
	−.16(1)	∗
17	−.15(−8)	
18	−.52(−18)	

shows that T_{11} is chosen as pivot if either

$$\chi_1 \le \frac{|T_{11}|}{\delta} \quad \text{or} \quad \chi_1 \le \sqrt{\frac{\chi_r |T_{11}|}{\delta}}. \tag{A.64}$$

Now, since χ_r is the maximum off-diagonal in some column of (26), we have $\chi_r = O(1)$, while since T_{11} comes from either the $(1,1)$ or $(2,2)$ block of (26), we have $|T_{11}| = O(\mu + \mathbf{u})$. Since $\delta \in (0,1)$ is fixed, we have from (A.64) that

$$\chi_1 = O(\mu^{1/2} + \mathbf{u}^{1/2}). \tag{A.65}$$

Since χ_1 is the magnitude of the largest off-diagonal in some row/column of (26), we have that χ_1 is the ∞-norm of some row or column of B. But (A.65) is incompatible with $B = O(1)$ and $\kappa(B) = O(1)$. Hence $|T_{11}|$ from the $(1,1)$ or $(2,2)$ blocks cannot be used as a pivot.

A similar argument holds when T_{rr} is chosen as a pivot, where T_{rr} is one of the small diagonals.

(iv) The pivot is 2×2 and involves at least one element from Λ. Since all the off-diagonals in (26) are $O(1)$, the quantities χ_i, $i = 1, 2, \ldots, \bar{n}$ are all $O(1)$. A 2×2 pivot with diagonal elements T_{11} and T_{rr} must have

$$|T_{11}| \leq \delta \chi_1, \qquad |T_{rr}| \leq \delta \chi_r,$$

which implies that T_{11} and T_{rr} are both $O(1)$. Since all the diagonals of Λ are $\Omega(\mu^{-1})$, they cannot be candidates for T_{11} and T_{rr}.

(v) The pivot is 2×2 and the pivot block is drawn either entirely from the $(1,1)$ block of (26) or entirely from the $(2,2)$ block. In this case, T_{1r} — the element for which $|T_{1r}| = \chi_1$ — is $O(\mu + \mathbf{u})$. Since T_{1r} has the largest magnitude in its column of (26) and since its column includes either a row or column of B, we have that one of the rows or columns of B is $O(\mu + \mathbf{u})$. As in (iii), we have a contradiction, since this estimate is incompatible with $B = \Omega(1)$ and $\kappa(B) = O(1)$.

This completes the proof of part (a).

We turn to (b), examining the effects of one step of elimination performed with pivot selection corresponding to the two cases (i) and (ii) of Theorem 6.1(a). For (i), suppose the (i,i) element of Λ is chosen as the pivot. After symmetric permutation of the canonical matrix to place the pivot in the $(1,1)$ position, we obtain

$$\left[\begin{array}{c|c} 1 & \\ \hline & \tilde{P} \end{array}\right] \left[\begin{array}{c|ccc} (\Lambda + O(\mu + \mathbf{u}))_{ii} & N_{i\cdot}^T & 0 & 0 \\ \hline N_{\cdot i} & 0 & B & \tilde{N} \\ 0 & B^T & 0 & 0 \\ 0 & \tilde{N}^T & 0 & \tilde{\Lambda} \end{array}\right] \left[\begin{array}{c|c} 1 & \\ \hline & \tilde{P}^T \end{array}\right] + O(\mu + \mathbf{u}),$$

where $\tilde{P}$ is some permutation matrix, $N_{\cdot i}$ denotes the ith column of N, $\tilde{N}$ is obtained from N by removing $N_{\cdot i}$, and $\tilde{\Lambda}$ is obtained from Λ by removing its ith row and column. Since $|(\Lambda + O(\mu + \mathbf{u}))_{ii}^{-1}| = O(\mu)$, the submatrix that remains after elimination is

$$\tilde{P} \begin{bmatrix} 0 & B & \tilde{N} \\ B^T & 0 & 0 \\ \tilde{N}^T & 0 & \tilde{\Lambda} \end{bmatrix} \tilde{P}^T - \tilde{P} \begin{bmatrix} N_{\cdot i} \\ 0 \\ 0 \end{bmatrix} \Lambda_{ii}^{-1} \begin{bmatrix} N_{i\cdot}^T & 0 & 0 \end{bmatrix} \tilde{P}^T + O(\mu + \mathbf{u})$$

$$= \tilde{P} \begin{bmatrix} 0 & B & \tilde{N} \\ B^T & 0 & 0 \\ \tilde{N}^T & 0 & \tilde{\Lambda} \end{bmatrix} \tilde{P}^T + O(\mu + \mathbf{u}). \tag{A.66}$$

It is easy to see that (A.66) is canonical, so our result is proved for case (i).

For case (ii), the proof is a little messier. Suppose the diagonals of the 2×2 pivot are the (i,i) element of E_1 and the (j,j) element of E_2. After symmetric rearrangement to put this pivot in the upper left corner, (26) becomes

$$\left[\begin{array}{c|c} I & \\ \hline & \hat{P} \end{array}\right] \left[\begin{array}{cc|ccc} 0 & B_{ij} & 0 & B_{i\cdot;j} & N_{i\cdot} \\ B_{ij} & 0 & B_{\cdot j;i}^T & 0 & 0 \\ \hline 0 & B_{\cdot j;i} & 0 & \hat{B} & \hat{N} \\ B_{i\cdot;j}^T & 0 & \hat{B}^T & 0 & 0 \\ N_{i\cdot}^T & 0 & \hat{N}^T & 0 & \Lambda \end{array}\right] \left[\begin{array}{c|c} I & \\ \hline & \hat{P}^T \end{array}\right] + O(\mu + \mathbf{u}),$$

where

- $\hat{P}$ is some permutation matrix,
- $N_{i\cdot}$ is the ith row of N,
- $\hat{N}$ is N with $N_{i\cdot}$ removed,
- $B_{i\cdot;j}$ is the ith row of B with its jth element removed,
- $B_{\cdot j;i}$ is the jth column of B with its ith element removed,
- $\hat{B}$ is B with its ith and jth column removed.

By the choice of B_{ij}, it is either the largest element in its row or the largest element in its column of B. From our assumptions on B, we deduce that $|B_{ij}| = \Omega(1)$. Denoting the pivot block by E, we have

$$E = B_{ij}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + O(\mu + \mathbf{u}), \qquad E^{-1} = \frac{1}{B_{ij}}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + O(\mu + \mathbf{u}). \tag{A.67}$$

Therefore, the elimination step yields the remaining matrix

$$\begin{aligned} &\hat{P}\begin{bmatrix} 0 & \hat{B} & \hat{N} \\ \hat{B}^T & 0 & 0 \\ \hat{N}^T & 0 & \Lambda \end{bmatrix}\hat{P}^T - \frac{1}{B_{ij}}\hat{P}\begin{bmatrix} 0 & B_{\cdot j;i} \\ B_{i\cdot;j}^T & 0 \\ N_{i\cdot}^T & 0 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 0 & B_{i\cdot;j} & N_{i\cdot} \\ B_{\cdot j;i}^T & 0 & 0 \end{bmatrix}\hat{P}^T \\ &\qquad + O(\mu + \mathbf{u}) \\ &= \hat{P}\begin{bmatrix} 0 & \hat{B} & \hat{N} \\ \hat{B}^T & 0 & 0 \\ \hat{N}^T & 0 & \Lambda \end{bmatrix}\hat{P}^T + O(\mu + \mathbf{u}), \end{aligned} \tag{A.68}$$

where

$$\bar{B} = \hat{B} - \frac{1}{B_{ij}}B_{\cdot j;i}B_{i\cdot;j}, \qquad \bar{N} = \hat{N} - \frac{1}{B_{ij}}B_{\cdot j;i}N_{i\cdot}.$$

It is obvious that (A.68) satisfies Definition 5.1, except possibly for the conditioning of the remaining matrix $\bar{B}$. This matrix is obtained by pivoting the (i, j) element of B to the $(1, 1)$ position and then doing one step of Gaussian elimination. In fact, we are doing partial pivoting since, as noted above, B_{ij} is the largest element in either its row or its column. Hence, the conditioning of the reduced submatrix $\bar{B}$ is unlikely to differ much from $\kappa(B)$, so it is reasonable to assert that $\kappa(\bar{B}) = O(1)$.

We have shown that our stated result holds for both cases (i) and (ii), so our proof of part (b) is complete.

For part (c), note that $C = O(1)$ whether the pivot block is 1×1 or 2×2. For 1×1 pivots, we have $|E| = \Omega(\mu^{-1})$ and $|E^{-1}| = \Omega(\mu)$. For 2×2 pivots, we have from (A.67) and $|B_{ij}| = \Omega(1)$ that $|E| = O(1)$ and $|E^{-1}| = O(1)$.

A.2. Proof of Theorem 6.2. We prove (a) of Theorem 6.2 by again excluding the other possible choice for a pivot.

(iii) The pivot is 2×2 and contains at least one element from Λ. In a degenerate canonical matrix, we have $\chi_i = O(\mu + \mathbf{u})$, $i = 1, 2, \ldots, \bar{n}$. A 2×2 pivot with diagonal elements T_{11} and T_{rr} must have

$$|T_{11}| \le \delta\chi_1, \qquad |T_{rr}| \le \delta\chi_r,$$

which implies that both diagonals are $O(\mu + \mathbf{u})$, so neither element can come from Λ.

In the case of either a 1×1 or 2×2 pivot made up of elements of size $O(\mu+\mathbf{u})$, we can use the standard argument about element growth in Bunch–Kaufman (that is, the argument that leads to (44)) to deduce the result (b). In the remaining case, where the pivot is a single diagonal element from Λ, we have in the notation of (42) that $|C| = O(\mu+\mathbf{u})$ and $|E| = \Omega(\mu^{-1})$. Hence, the update to the remaining submatrix is bounded by

$$|C||E^{-1}||C|^T = O(\mu(\mu+\mathbf{u})^2),$$

which certainly has size $O(\mu+\mathbf{u})$.

A.3. Proof of Lemma 6.4.

Proof. In floating-point arithmetic, the LU factorization of (50) yields the following approximate LU factors:

$$\text{(A.69)}\qquad \begin{bmatrix} 1 & 0 \\ E_{11}/E_{12}+\delta_1 & 1 \end{bmatrix}, \qquad \begin{bmatrix} E_{12} & E_{22} \\ 0 & E_{12}-E_{11}E_{22}/E_{12}+\delta_2 \end{bmatrix},$$

where

$$\delta_1 = \left|\frac{E_{11}}{E_{12}}\right| O(\mathbf{u}), \qquad \delta_2 = |E_{12}|O(\mathbf{u}) + |E_{11}E_{22}/E_{12}|\, O(\mathbf{u}).$$

It is well known that for triangular substitution applied to any triangular system $Uz = h$, the computed solution $\hat{z}$ satisfies $(U+\Delta_U)\hat{z} = h$, where $|E_U| = |U|O(\mathbf{u})$. By applying this observation to each of the matrices in (A.69), we find that the computed solution $\hat{y}$ of (50) satisfies

$$\text{(A.70)}\qquad \begin{bmatrix} 1 & 0 \\ E_{11}/E_{12}+\delta_3 & 1 \end{bmatrix} \begin{bmatrix} E_{12}+\delta_4 & E_{22}+\delta_5 \\ 0 & E_{12}-E_{11}E_{22}/E_{12}+\delta_6 \end{bmatrix} \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} = \begin{bmatrix} g_2 \\ g_1 \end{bmatrix},$$

where

$$\begin{aligned}
\delta_3 &= \delta_1 + |E_{11}/E_{12}|\, O(\mathbf{u}) = |E_{11}/E_{12}|\, O(\mathbf{u}),\\
\delta_4 &= |E_{12}|O(\mathbf{u}),\\
\delta_5 &= |E_{22}|O(\mathbf{u}),\\
\delta_6 &= \delta_2 + (|E_{12}| + |E_{11}E_{22}/E_{12}|)\, O(\mathbf{u}) = |E_{12}|O(\mathbf{u}) + |E_{11}E_{22}/E_{12}|\, O(\mathbf{u}).
\end{aligned}$$

By multiplying out the coefficient matrix in (A.70), we obtain

$$\text{(A.71)}\qquad \begin{bmatrix} E_{12}+\delta_4 & E_{22}+\delta_5 \\ E_{11}+\delta_7 & E_{12}+\delta_8 \end{bmatrix},$$

where

$$\begin{aligned}
\delta_7 &= |E_{12}|\delta_3 + |E_{11}/E_{12}|\, \delta_4 = |E_{11}|O(\mathbf{u}),\\
\delta_8 &= |E_{11}/E_{12}|\, \delta_5 + |E_{22}|\delta_3 + \delta_+ (|E_{12}| + |E_{11}E_{22}/E_{12}|)\, O(\mathbf{u})\\
&= |E_{11}|O(\mathbf{u}) + |E_{11}E_{22}/E_{12}|\, O(\mathbf{u}) + |E_{12}|O(\mathbf{u})\\
&= |E_{12}|O(\mathbf{u}).
\end{aligned}$$

(The last equality follows from (49).) Hence, (A.71) can be written as

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} (E + \Delta_E),$$

where Δ_E satisfies the bound (51). ∎

A.4. Proof of Theorem 8.1.

Proof. We start by proving the analog of Theorem 6.1(a). As in the earlier proof, we systematically exclude the three other possible choices of pivots.

(iii) The pivot is 1×1 and is a diagonal element from either the $(1,1)$ or $(2,2)$ blocks of (26). Then this pivot (T_{ii}, say) will be $O(\mu + \mathbf{u})$. According to the stability criterion (56) we then have $\chi_i = O(\mu + \mathbf{u})$, which implies that one of the rows or columns of B is $O(\mu + \mathbf{u})$. However, this estimate is incompatible with $B = \Omega(1)$ and $\kappa(B) = O(1)$, so this kind of pivot cannot occur.

(iv) The pivot is 2×2 and involves at least one diagonal element from Λ. First, we show that we cannot have both diagonals from Λ. If this were the case, then at least one of these diagonals (T_{ii}, say) would have been considered as a 1×1 pivot at an earlier point in the algorithm. But if it was considered, it would have been accepted since

$$|T_{ii}^{-1}|\chi_i = O(\mu)O(1) = O(\mu) \leq 2/\delta$$

for sufficiently small μ. Hence, at most one of the diagonals is from Λ. Without loss of generality, suppose in (57) that T_{ii} is from Λ while the remaining diagonal T_{jj} is $O(\mu + \mathbf{u})$. In fact, we have

$$T_{ii} = \Omega(\mu^{-1}), \qquad T_{jj} = O(\mu + \mathbf{u}), \qquad T_{ij} = O(1),$$

and so

$$\left| \begin{bmatrix} T_{ii} & T_{ij} \\ T_{ij} & T_{jj} \end{bmatrix}^{-1} \right| = \frac{1}{|T_{ii}T_{jj} - T_{ij}^2|} \left| \begin{bmatrix} T_{jj} & -T_{ij} \\ -T_{ij} & T_{ii} \end{bmatrix} \right|.$$

Hence, from (57), we have

$$\left| \begin{bmatrix} T_{jj} & -T_{ij} \\ -T_{ij} & T_{ii} \end{bmatrix} \right| \begin{bmatrix} \chi_i \\ \chi_j \end{bmatrix} \leq |T_{ii}T_{jj} - T_{ij}^2| \begin{bmatrix} 1/\delta \\ 1/\delta \end{bmatrix} = \begin{bmatrix} O(1) \\ O(1) \end{bmatrix}.$$

From the second row of this inequality, we have

$$\chi_j \leq \frac{1}{|T_{ii}|} O(1) = O(\mu).$$

But χ_j is the ∞-norm of one of the rows or columns of B, so this estimate contradicts our assumptions on B. Hence, this type of pivot cannot occur.

(v) The pivot is 2×2 and the pivot block E is drawn either entirely from the $(1,1)$ block of (26) or entirely from the $(2,2)$ block. In this case, all elements of E are $O(\mu + \mathbf{u})$. From (57), we have as above that

$$\left| \begin{bmatrix} T_{jj} & -T_{ij} \\ -T_{ij} & T_{ii} \end{bmatrix} \right| \begin{bmatrix} \chi_i \\ \chi_j \end{bmatrix} \leq |T_{ii}T_{jj} - T_{ij}^2| O(1).$$

Taking the second row of this relation, we obtain

$$|T_{ij}|\chi_i + |T_{ii}|\chi_j \leq |T_{ii}T_{jj} - T_{ij}^2|O(1) \leq \left(|T_{ii}T_{jj}| + |T_{ij}|^2\right) O(1), \tag{A.72}$$

where, by definition, χ_i and χ_j are both nonnegative. Consider two cases. When $|T_{ij}|^2 \geq |T_{ii}T_{jj}|$ we have from (A.72) that

$$|T_{ij}|\chi_i \leq |T_{ij}|^2 O(1) \implies \chi_i = O(|T_{ij}|) = O(\mu + \mathbf{u}).$$

For the reasons outlined earlier, the assumptions on B are inconsistent with this bound on χ_i, so this case cannot hold. For the other case $|T_{ij}|^2 < |T_{ii}T_{jj}|$, we have

$$|T_{ii}|\chi_j \leq |T_{ii}T_{jj}|O(1) \implies \chi_j = O(|T_{jj}|) = O(\mu + \mathbf{u}),$$

which is also disallowed by our assumptions. Hence, pivots of this type cannot occur.

The proof of the remaining parts (b) and (c) of Theorem 6.1 is identical in this case.

Turning now to the case of a degenerate canonical matrix and the analog of Theorem 6.2, we start by showing that no 2×2 pivots may contain diagonal elements from Λ.

Note that for a degenerate matrix, the off-diagonals, and hence the quantities χ_i, all have size $O(\mu + \mathbf{u})$. If the pivot is a 2×2 block in which both diagonals are from Λ, then one of them (T_{ii}, say) must have been considered previously as a 1×1 pivot. But if it was considered, it would have been accepted since

$$|T_{ii}^{-1}|\chi_i = O(\mu)O(\mu + \mathbf{u}) \leq 2/\delta.$$

Hence, this type of pivot cannot occur.

If just one of the diagonals is from Λ, this diagonal element (T_{jj}, say) must not have been considered earlier as a 1×1 pivot, since then it would have been accepted for the reason described above. Hence, the other pivot T_{ii}, which has size $O(\mu + \mathbf{u})$, must have been considered as a 1×1 pivot and rejected. Because of (56), T_{ii} must satisfy

$$|T_{ii}| < \frac{\delta}{2}\chi_i. \tag{A.73}$$

On the other hand, since the 2×2 pivot is accepted, we must have

$$\left| \begin{bmatrix} T_{jj} & -T_{ij} \\ -T_{ij} & T_{ii} \end{bmatrix} \right| \begin{bmatrix} \chi_i \\ \chi_j \end{bmatrix} \leq |T_{ii}T_{jj} - T_{ij}^2| \begin{bmatrix} 1/\delta \\ 1/\delta \end{bmatrix}. \tag{A.74}$$

Consider first the case of $T_{ij}^2 \geq |T_{ii}T_{jj}|$. Then from the first block row in (A.74), this inequality implies that

$$|T_{jj}|\chi_i \leq |T_{ii}T_{jj} - T_{ij}^2|\frac{1}{\delta} \leq 2T_{ij}^2\frac{1}{\delta}.$$

Since $|T_{ij}| \leq \chi_i$, we have

$$|T_{jj}| \leq 2|T_{ij}|\frac{1}{\delta} = O(\mu + \mathbf{u}),$$

which contradicts our assumption that T_{jj} has size $\Omega(\mu^{-1})$. The remaining case has $T_{ij}^2 < |T_{ii}T_{jj}|$. From (A.74) and (A.73), we have

$$|T_{jj}|\chi_i \leq 2|T_{ii}T_{jj}|\frac{1}{\delta} < 2|T_{jj}|\frac{1}{\delta}\frac{\delta}{2}\chi_i = |T_{jj}|\chi_i,$$

which is a contradiction. Hence this kind of pivot — in which exactly one of the diagonals comes from Λ — cannot occur either, and we are done.

For the analog of part (b) of Theorem 6.2, we have from (56) and (57) and the definition of C and E in (42) that

$$|E^{-1}C^T| \leq |E^{-1}||C^T| = O(1/\delta) = O(1).$$

Hence, the update matrix $CE^{-1}C^T$ is bounded as follows:

$$|CE^{-1}C^T| = \|C\|O(1) = O(\mu + \mathbf{u}),$$

giving the result. □

Acknowledgments. I thank the editor, Linda Kaufman, for the care she took with this paper, and two anonymous referees for their meticulous reports which improved both presentation and content. I also thank John Lewis for interesting discussions about the Bunch–Kaufman factorization during his visit to Argonne in April 1995.

REFERENCES

[1] C. Ashcraft, R. L. Grimes, and J. G. Lewis, *Accurate symmetric indefinite linear equation solvers*, manuscript.

[2] J. Bunch and L. Kaufman, *Some stable methods for calculating inertia and solving symmetric linear systems*, Math. Comput., 31 (1977), pp. 163–179.

[3] I. S. Duff, *The Solution of Augmented Systems*, Technical report RAL-93-084, Rutherford Appleton Laboratory, Oxon, U. K., 1993.

[4] A. Forsgren, P. Gill, and J. Shinnerl, *Stability of Symmetric Ill-Conditioned Systems Arising in Interior Methods for Constrained Optimization*, Report TRITA-MAT-1994-24, Royal Institute of Technology, 1994.

[5] R. Fourer and S. Mehrotra, *Solving symmetric indefinite systems in an interior-point method for linear programming*, Math. Programming, 62 (1993), pp. 15–39.

[6] A. J. Goldman and A. W. Tucker, *Theory of linear programming*, in Linear Equalities and Related Systems, H. W. Kuhn and A. W. Tucker, eds., Princeton University Press, Princeton, NJ, 1956, pp. 53–97.

[7] C. Gonzaga, *Path-following methods in linear programming*, SIAM Rev., 34 (1991), pp. 167–224.

[8] O. Güler and Y. Ye, *Convergence behavior of interior-point algorithms*, Math. Programming, 60 (1993), pp. 215–228.

[9] M. Kojima, S. Mizuno, and A. Yoshise, *An $O(\sqrt{n}L)$ iteration potential reduction algorithm for linear complementarity problems*, Math. Programming, 50 (1991), pp. 331–342.

[10] I. J. Lustig, R. E. Marsten, and D. F. Shanno, *Computational experience with a primal-dual interior point method for linear programming*, Linear Algebra Appl., 152 (1991), pp. 191–222.

[11] I. J. Lustig, R. E. Marsten, and D. F. Shanno, *Computational experience with a globally convergent primal-dual predictor-corrector algorithm for linear programming*, Math. Programming, 66 (1994), pp. 123–135.

[12] 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.

[13] N. Megiddo, *On finding primal- and dual-optimal bases*, ORSA J. Comput., 3 (1991), pp. 63–65.

[14] S. Mehrotra, *On the implementation of a primal-dual interior point method*, SIAM J. Optim., 2 (1992), pp. 575–601.

[15] S. Mizuno, M. Todd, and Y. Ye, *On adaptive step primal-dual interior-point algorithms for linear programming*, Math. Oper. Res., 18 (1993), pp. 964–981.
[16] D. B. Ponceleón, *Barrier Methods for Large-scale Quadratic Programming*, Ph.D. thesis, Stanford University, Stanford, CA, 1990.
[17] M. Todd, *Potential Reduction Methods in Mathematical Programming*, Technical report 1112, Cornell University, Ithaca, NY, 1995.
[18] R. J. Vanderbei, *LOQO User's Manual*, Technical report SOR 92-5, Program in Statistics and Operations Research, Princeton University, Princeton, NJ, 1992.
[19] S. Vavasis, *Stable numerical algorithms for equilibrium systems*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1108–1131.
[20] S. J. Wright, *A path-following infeasible-interior-point algorithm for linear complementarity problems*, Optimization Methods Software, 2 (1993), pp. 79–106.
[21] S. J. Wright, *A path-following interior-point algorithm for linear and quadratic optimization problems*, Ann. Oper. Res., 62 (1996), pp. 103–130.
[22] S. J. Wright, *Stability of linear equations solvers in interior-point methods*, SIAM J. Matrix Anal. Appl., 16 (1994), pp. 1287–1307.
[23] X. Xu, P. Hung, and Y. Ye, *A simplified homogeneous and self-dual linear programming algorithm and its implementation*, Ann. Oper. Res., 62 (1996), pp. 151–172.
[24] Y. Ye, *On the finite convergence of interior-point algorithms for linear programming*, Math. Programming, 57 (1992), pp. 325–336.
[25] Y. Zhang, *On the convergence of a class of infeasible-interior-point methods for the horizontal linear complementarity problem*, SIAM J. Optim., 4 (1994), pp. 208–227.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 223–248, January 1997

016

ON EPISODIC QUEUES*

QI-MING HE† AND MARCEL F. NEUTS‡

Abstract. For a queueing system with a low traffic intensity, the expected number of customers in the queue is small. However, with a bursty input process, long queues may build up in a short time. In this paper, we study light traffic queueing systems with bursty input processes. We study the distributions of queue lengths, waiting times, busy and active periods, and their corresponding expansions when the traffic intensity is small. Special attention goes to some conditional distributions of queue lengths, waiting times, and system-active periods. The expansions given in this paper provide a potential asymptotic approach to the computation of various descriptors of queueing systems. The coefficients of those expansions reflect some important features of episodic queues. We also report numerical results which give a graphic view of our approximations and of the effect of the burstiness of the input and service processes on the queues.

Key words. queueing theory, Markov arrival processes, burstiness, quasi-birth-and-death processes, algorithmic probability

AMS subject classification. 60K25

PII. S0895479895281472

1. Introduction.

In an episodic queue, (brief) periods of activity alternate with (longer) periods during which there is no arrival. Episodic behavior may occur even in queues with a very low traffic intensity. During the periods of activity, substantial queue lengths may build up. Behavior of that type, which is common in queues with "bursty" arrival processes, is inadequately measured by the traditional descriptors such as the lower-order moments of the steady-state queue length and waiting time distributions. For example, in an episodic queue, the queue length in the periods of activity might be dramatically different from the queue length in the periods during which there is no arrival. Hence the expected queue length may not reflect the average queue lengths in either type of period.

Episodic queues can be seen in the banking industry. Consider a special type of transaction for which a high volume of data arrives during certain periods of each day (or month) while there is little work to be done for the rest of the day (or month). In such systems, it is more useful to know what is happening during the peak periods of transaction than the overall averages.

In this paper, we shall study a $MAP/MAP/1$ queue under particular episodic conditions. Various conditional distributions and their asymptotic expansions will be discussed and, through numerical examples and graphs, we shall show how they describe useful features of the queue. We focus on the $MAP/MAP/1$ queue because of its tractability by matrix-analytic methods. Also, the MAP is a convenient and versatile tool for representing point processes with varying arrival rates. For a brief introduction to the MAP (Markovian arrival process), see Appendix A. For discussions of the properties of the Markovian arrival process and its use in queueing models, we refer, e.g., to Lucantoni [4] or Neuts [5, 9].

* Received by the editors February 13, 1995; accepted for publication (in revised form) by Y. Genin February 6, 1996. This research was supported in part by the K. C. Wang Education Foundation, Bell Communications Research, and National Science Foundation grant DDM-8915235.
http://www.siam.org/journals/simax/18-1/28147.html

† Department of Management Sciences, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

‡ Department of Systems and Industrial Engineering, The University of Arizona, Tucson, AZ 85721.

The approach in this paper is somewhat related to light traffic approximations, as treated in Reiman and Simon [11, 12]. More recently, Blaszczyszyn, Frey, and Schmidt [1] considered a light traffic approximation for Markov-modulated multiserver queues. In this paper, factorial moment expansions are applied to derive approximation formulas for various performance measures. Another paper by Van Den Hout and Blanc [13] applied the power series algorithm to study queueing systems with a *MAP*. In our research, the idea of series expansion is used as a mathematical tool in the analysis. The objective of this study is to show how an episodic queue performs. For this purpose, various conditional performance measures are defined and analyzed.

The organization of the paper is as follows. After this introduction, we describe a class of MAPs in which bursts of arrivals are separated by long intervals during which no arrivals occur (section 2.) In section 3, the steady-state distribution of the queue length is derived along with asymptotic expansions. These are used in sections 4 and 5 to obtain expressions and asymptotic expansions for various conditional queue lengths and waiting time distributions. Section 6 is devoted to the study of the busy and active periods of the queue. These reflect the interesting behavior of episodic queues. In the final section, we present numerical examples and graphs to illustrate the use of our descriptors of the queue and to show the quality of the approximations inherent in the asymptotic expansions.

2. Modeling. We consider a single server queueing system whose input and service processes are independent MAPs with coefficient matrices $(D_0(r), D_1)$ of order m_1 and (C_0, C_1) of order m_2, respectively. In order to specify the behavior at the boundary, we assume that, at the beginning of each busy period, the service process is restarted by selecting the service phase according to independent multinomial trials with the probability vector θ.

We consider parameter matrices for the input process of the following form:

$$D_0(r) = \begin{pmatrix} D_{01} & D_{02} \\ rD_{03} & rD_{04} + D_{05} \end{pmatrix} \text{ and } D_1 = \begin{pmatrix} D_{11} & D_{12} \\ 0 & 0 \end{pmatrix}, \tag{1}$$

where D_{01} and D_{11} are $m_{11} \times m_{11}$ matrices and D_{04} and D_{05} are $m_{12} \times m_{12}$ matrices. $D(r) = D_0(r) + D_1$ is an irreducible infinitesimal generator. D_{05} is either a zero matrix or an irreducible infinitesimal generator. r is a nonnegative parameter.

Our queueing system is denoted by $MAP(r)/MAP/1$ to emphasize its dependence on the parameter r. In what follows, we shall consider small values of r.

We see that there are no arrivals during the m_{12} phases corresponding to the second row of the matrix $D(r)$. For these, there are possibly long intervals without arrivals. By specifying the various submatrices of the parameter matrices $D_0(r)$ and D_1, we can model a variety of patterns of burstiness in the periods with arrivals and of durations of the periods without.

The following terminology is useful in the discussion of the episodic queue. Since arrivals occur only in the first m_{11} phases of $D(r)$, we call the input process *input-active* when it is in one of these phases and *input-inactive* otherwise. We call the queueing system *system-active* if the input process is active or if the server is busy and *system-inactive* otherwise.

We denote by $\theta_1(r)$ the stationary probability vector of $D(r)$. The vector $\theta_1(r)$ is partitioned into $[\theta_{11}(r), \theta_{12}(r)]$, where $\theta_{11}(r)$ has dimension m_{11}. $\theta_{11}(r)$ and $\theta_{12}(r)$ satisfy

$$0 = \theta_{11}(r)(D_{01} + D_{11}) + r\theta_{12}(r)D_{03},$$

$$0 = \theta_{11}(r)(D_{02} + D_{12}) + \theta_{12}(r)(rD_{04} + D_{05}). \tag{2}$$

The arrival rate of the input MAP is $\lambda^*(r) = \theta_1(r)D_1\mathbf{e}$, where $\mathbf{e}$ is the column vector with all components 1. When the input process is stationary, an arbitrary input-inactive period has a phase type (PH)-distribution (see Neuts [6]) whose representation consists of the probability vector $[\theta_{11}(r)(D_{02} + D_{12})\mathbf{e}]^{-1}\theta_{11}(r)(D_{02} + D_{12})$ and of the matrix $rD_{04} + D_{05}$.

$\omega(r)$, the mean of an arbitrary input-inactive period, is given by

$$\omega(r) = \frac{\theta_{12}(r)\mathbf{e}}{r \cdot \theta_{12}(r)D_{03}\mathbf{e}}.$$

Next we study relations between $\lambda^*(r)$, $\omega(r)$, and r.

LEMMA 2.1. *$\theta_1(r)$ is an analytic vector function of r and*

$$\lim_{r\to 0} \theta_1(r) = \begin{cases} (0, \theta_3) & \text{if } D_{05} = 0, \\ \\ (0, \theta_4) & \text{if } D_{05} \neq 0 \text{ and irreducible}, \end{cases} \tag{3}$$

where θ_3 is the stationary probability vector of the infinitesimal generator $Q_{22} = D_{04} - D_{03}(D_{01} + D_{11})^{-1}(D_{02} + D_{12})$ and θ_4 is the stationary probability vector of D_{05}. $\lambda^*(r)$ and $\omega(r)$ are analytic functions of r. The first two terms in the expansion

$$\lambda^*(r) = r\lambda_1^* + r^2\lambda_2^* + o(r^2), \tag{4}$$

the coefficients λ_1^* and λ_2^*, as well as the quantity $\omega(r)$, are given by,
if $D_{05} = 0$,

$$\lambda_1^* = -\theta_3 D_{03}(D_{01} + D_{11})^{-1}(D_{11}\mathbf{e} + D_{12}\mathbf{e}),$$

$$\lambda_2^* = \lambda_1^*\theta_3 D_{03}(D_{01} + D_{11})^{-1}\mathbf{e},$$

$$\omega(r) = \frac{1}{r} \cdot \frac{1}{\theta_3 D_{03}\mathbf{e}};$$

if $D_{05} \neq 0$,

$$\lambda_1^* = -\theta_4 D_{03}(D_{01} + D_{11})^{-1}(D_{11}\mathbf{e} + D_{12}\mathbf{e}),$$

$$\lambda_2^* = -\theta_{12}' D_{03}(D_{01} + D_{11})^{-1}(D_{11}\mathbf{e} + D_{12}\mathbf{e}),$$

$$\omega(r) = \frac{1}{r}\left\{\frac{1}{\theta_4 D_{03}\mathbf{e}} + r\left[\frac{\theta_{12}'\mathbf{e}}{\theta_4 D_{03}\mathbf{e}} - \frac{\theta_{12}' D_{03}\mathbf{e}}{(\theta_4 D_{03}\mathbf{e})^2}\right] + o(r)\right\},$$

where $\theta_{12}' = \theta_4 Q_{22}(\mathbf{e}\theta_4 - D_{05})^{-1} + \theta_4 D_{03}(D_{01} + D_{11})^{-1}\mathbf{e}\theta_4$.

Proof. See Appendix B. □

Let $C = C_0 + C_1$ be irreducible. The service rate is $\mu^* = \theta_2 C_1 \mathbf{e}$, where θ_2 is the stationary probability vector of C. The traffic intensity of the queue is therefore given by

$$\rho(r) = \frac{\lambda^*(r)}{\mu^*},$$

and the queue is stable iff $\rho(r) < 1$. From Lemma 2.1, the following result is obvious.

COROLLARY 2.2.

$$\rho(r) \to 0 \Leftrightarrow \lambda^*(r) \to 0 \Leftrightarrow r \to 0 \Leftrightarrow \omega(r) \to \infty. \qquad \square$$

When $r \to 0$, the durations of the input-inactive periods increase. Any two successive system-active periods are therefore separated by a long system-inactive period. Furthermore, with a small value of r, the traffic intensity is small. That parameter sheds no light on the substantial excursions during the active periods.

Example 1. If the arrivals to a single server queue with *MAP* service consist of all the customers blocked from an *M/M/n/n* queue, the input process is a *MAP* with coefficient matrices

$$D_0 = \begin{pmatrix} -\lambda & \lambda & & & & \\ \mu & -\mu-\lambda & \lambda & & & \\ & \ddots & \ddots & \ddots & & \\ & & n\mu-\mu & -n\mu-\mu-\lambda & \lambda & \\ & & & n\mu & -n\mu-\lambda & \end{pmatrix},$$

$$D_1 = \begin{pmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & \lambda \end{pmatrix},$$

where λ and μ are, respectively, the arrival and service rate for the *M/M/n/n* queue. When the number of customers in the *M/M/n/n* queue is fewer than n, there are no arrivals to the system.

Example 2. Consider the $MMPP(r)/MAP/1$ queue (see Lucantoni [4]), where the input process is a Markov-modulated Poisson process. We assume that the $MMPP(r)$ is given by (1) with

$$D_{11} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{m_{11}} \end{pmatrix} \text{ and } \quad D_{12} = 0.$$

The input process is active in the first m_{11} states and there are no arrivals on the remaining m_{12} states.

3. The queue length at an arbitrary time. Before discussing the conditional queues in section 4, we give a detailed analysis of the queue length at an arbitrary time.

We denote by $q(t)$ the queue length, by $J_1(t)$ the phase of the input MAP, and by $J_2(t)$ the phase of the service process at time t. $J_2(t)$, the phase of the service MAP, is defined only when the server is busy. (We recall the assumption that the service phase at the beginning of each busy period is selected according to the probability vector θ.) Then $(q(t), J_1(t), J_2(t))$ is an irreducible continuous-time Markov chain with infinitesimal generator

$$Q(r) = \begin{pmatrix} D_0(r) & D_1 \otimes \theta & & & \\ I \otimes C_1\mathbf{e} & D_0(r) \oplus C_0 & D_1 \otimes I & & \\ & I \otimes C_1 & D_0(r) \oplus C_0 & D_1 \otimes I & \\ & & \ddots & \ddots & \ddots \end{pmatrix},$$

where $\otimes$ and $\oplus$ denote, respectively, the *Kronecker product* and the *Kronecker sum* of matrices. We denote by $\mathbf{X}(r) = (\mathbf{x}_0(r), \mathbf{x}_1(r), \ldots)$ the stationary probability vector of the Markov process $(q(t), J_1(t), J_2(t))$. By the classical matrix-geometric theorem in Neuts [6],

$$\mathbf{x}_n(r) = \mathbf{x}_0(r)(I \otimes \theta)[R(r)]^n, \quad n \geq 1, \tag{5}$$

and $\mathbf{x}_0(r)$ satisfies equations

$$\begin{aligned} 0 &= \mathbf{x}_0(r)[D_0(r) + (I \otimes \theta)R(r)(I \otimes (C_1\mathbf{e}))], \\ 1 &= \mathbf{x}_0(r)(I \otimes \theta)(I - R(r))^{-1}\mathbf{e}. \end{aligned} \tag{6}$$

Equation (6) is discussed in detail in Appendix C. The rate matrix $R(r)$ is the minimal nonnegative solution to the equation

$$0 = R^2(r)(I \otimes C_1) + R(r)[D_0(r) \oplus C_0] + D_1 \otimes I. \tag{7}$$

Since D_1 is of the form (1), $R(r)$ has the special structure

$$R(r) = \begin{pmatrix} R_1(r) & R_2(r) \\ 0 & 0 \end{pmatrix},$$

where $R_1(r)$ and $R_2(r)$ are $(m_{11}m_2) \times (m_{11}m_2)$ and $(m_{11}m_2) \times (m_{12}m_2)$ nonnegative matrices, respectively.

In what follows, we shall expand $\mathbf{X}(r)$ in powers of r. By (5) and (6), the expansion of $\mathbf{X}(r)$ is determined by the expansions of $R(r)$ and $\mathbf{x}_0(r)$. So, we first establish the necessary differentiability properties of $R^{(n)}(r)$ and $\mathbf{x}_0(r)$ and calculate the expansion of $R(r)$.

LEMMA 3.1. *The derivative $R^{(n)}(r)$ is differentiable in r for all $n \geq 0$. $\lim_{r\to 0} R(r) = R$ exists and is finite. R is the minimal nonnegative solution to (7) with $r = 0$. $\mathrm{sp}(R)$, the largest eigenvalue of R, is less then 1.*

Proof. That $R(r)$ is differentiable in r is obtained from Theorem 2.2 in He [2].

By the diagonal method, it is easy to see that the limit of $R(r)$ as $r \to 0$ exists and is the minimal nonnegative solution to (7) with $r = 0$. $R_1 = \lim_{r\to 0} R_1(r)$ is the minimal nonnegative solution to the following equation:

$$0 = R_1^2(I \otimes C_1) + R_1(D_{01} \oplus C_0) + D_{11} \otimes I. \tag{8}$$

Clearly, $\mathrm{sp}(R_1) < 1$ implies that $\mathrm{sp}(R) < 1$. □

The set of states $\{(i, j, k), 1 \le j \le m_1, 1 \le k \le m_2\}$ is called *level* i for $i \ge 1$. Likewise, the set $\{(0, j), 1 \le j \le m_1\}$ is called *level* 0. We divide level 0 into two sublevels 0_1, consisting of the first m_{11} states of level 0, and 0_2, consisting of the remaining states. We accordingly partition $\mathbf{x}_0(r)$ into $(\mathbf{x}_{01}(r), \mathbf{x}_{02}(r))$.

LEMMA 3.2. *The derivative* $\mathbf{x}_0^{(n)}(r)$ *is differentiable in* r *for all* $n \ge 0$ *and* $\mathbf{x}_0(r)$ *satisfies*

$$(\mathbf{x}_0(r) \otimes \theta)(I - R(r))^{-1}(I \otimes \mathbf{e}) = \theta_1(r). \tag{9}$$

$\lim_{r\to 0} \mathbf{x}_0(r) = \lim_{r\to 0} \theta_1(r)$.

Proof. From equation (6) and Lemma 3.1, it follows that $\mathbf{x}_0^{(n)}(r)$ is differentiable in r for $n \ge 0$. Postmultiplying by $I \otimes \mathbf{e}$ on both sides of equation (7), we have

$$R(r)(I \otimes (C_1\mathbf{e})) = (I - R(r))^{-1}[R(r)(D_0(r) \otimes \mathbf{e}) + D_1 \otimes \mathbf{e}].$$

Substitution into equation (6) yields

$$(\mathbf{x}_0(r) \otimes \theta)(I - R(r))^{-1}(I \otimes \mathbf{e})D(r) = 0.$$

Equation (9) is obtained by (6). From equation (6), we have

$$\mathbf{x}_{01}(r) = r\mathbf{x}_{02}(r)D_{03}U_{11}^{-1}(r),$$

where $U_{11}(r) = -[D_{01}+(I\otimes\theta)R_1(r)(I\otimes(C_1\mathbf{e}))]$. Equation (7) implies that $-U_{11}(r)\mathbf{e} < 0$. By the irreducibility of $D(r)$ and the Markov process $Q(r)$, $U_{11}(r)$ is invertible (including at $r = 0$). So, as $r \to 0$, the limit of $\mathbf{x}_{01}(r)$ is 0. We rewrite (9) as

$$\mathbf{x}_0(r) + (\mathbf{x}_{01}(r) \otimes \theta, 0)R(r)(I - R(r))^{-1}(I \otimes \mathbf{e}) = \theta_1(r).$$

The limit of $\mathbf{x}_0(r)$ is therefore the same as $\theta_1(r)$. □

We now assume that $R(r)$ has the expansion

$$R(r) = R + rR' + r^2R'' + o(r^2).$$

Then (7) yields

$$0 = (RR' + R'R)(I \otimes C_1) + R'(D_0 \oplus C_0) + R(\bar{D}_0 \otimes I), \tag{10}$$

$$0 = (RR'' + (R')^2 + R''R)(I \otimes C_1) + R''(D_0 \oplus C_0) + R'(\bar{D}_0 \otimes I), \tag{11}$$

where $\bar{D}_0 = \begin{pmatrix} 0 & 0 \\ D_{03} & D_{04} \end{pmatrix}$. To solve R' and R'' from (10) and (11), we introduce the following transform.

For an $m \times n$ matrix A, $\phi(A)$ is the *direct sum* of the rows of A (see Neuts [6, section 3.9]). We shall, for brevity, call this construction the ϕ-transform. For given

m, n and a vector $\mathbf{u}$ of order mn, the inverse transform of ϕ at $\mathbf{u}$ is the $m \times n$ matrix A, for which $\phi(A) = \mathbf{u}$.

THEOREM 3.3. *The ϕ-transforms of R' and R'' are given explicitly by*

$$\phi(R') = \phi(R(\bar{D}_0 \otimes I))[I \otimes Z(0) - R^T \otimes (I \otimes C_1)]^{-1},$$

$$\phi(R'') = \phi([R'(\bar{D}_0 \otimes I) + (R')^2(I \otimes C_1)])[I \otimes Z(0) - R^T \otimes (I \otimes C_1)]^{-1},$$

where $Z(r) = -[D_0(r) \oplus C_0 + R(r)(I \otimes C_1)]$ and the superscript T denotes the transpose. The inverse transform of ϕ gives R' and R''.

Proof. From (10),

$$R'Z(0) - RR'(I \otimes C_1) = R(\bar{D}_0 \otimes I). \tag{12}$$

Taking ϕ-transforms on both sides of this equation gives the formula for R'. The invertibility of the matrix $I \otimes Z(0) - R^T \otimes (I \otimes C_1)$ follows from the fact that the largest eigenvalue of R is less than 1.

A similar argument leads to the result for R''. □

Note 1. We assume that $R_1(r)$ and $R_2(r)$ have the expansions $R_i(r) = R_i + rR_i' + r^2R_i'' + o(r^2)$, $i = 1, 2$. The matrices R_1, R_2, R_1', R_2', R_1'', and R_2'' can be obtained either from R, R', and R'', or by solving recursive equations. For example, R_1 is the minimal nonnegative solution to (8). R_2 satisfies the equation

$$0 = R_1R_2(I \otimes C_1) + R_1(D_{02} \otimes I) + R_2(D_{05} \oplus C_0) + D_{12} \otimes I.$$

When R_1 is obtained, R_2, R_1', R_2', R_1'', and R_2'' can be obtained explicitly by applying the ϕ-transform. So, the computation of R_1 is essential to all.

THEOREM 3.4. *The vector $\mathbf{X}^{(n)}(r)$ is differentiable in r for $n \geq 0$ and $\mathbf{X}(r)$'s expansion at $r = 0$ is given by*

$$\mathbf{x}_{01}(r) = r\mathbf{x}_{01}' + r^2\mathbf{x}_{01}'' + o(r^2),$$

$$\mathbf{x}_{02}(r) = \mathbf{x}_{02} + r\mathbf{x}_{02}' + r^2\mathbf{x}_{02}'' + o(r^2),$$

$$\mathbf{x}_n(r) = r(\mathbf{x}_{01}' \otimes \theta, 0)R^n + r^2\Bigg[(\mathbf{x}_{01}'' \otimes \theta, 0)R^n + (\mathbf{x}_{01}' \otimes \theta, 0)\sum_{i=0}^{n-1} R^iR'R^{n-i-1}\Bigg] + o(r^2), \quad n \geq 1, \tag{13}$$

where $\mathbf{x}_{02} = \theta_3$ when $D_{05} = 0$ and $\mathbf{x}_{02} = \theta_4$ when $D_{05} \neq 0$ and D_{05} is irreducible. $-\mathbf{x}_{02}'\mathbf{e} > \mathbf{x}_{01}'\mathbf{e} \geq 0$. $\{\mathbf{x}_{01}', \mathbf{x}_{02}', \mathbf{x}_{01}'', \mathbf{x}_{02}''\}$ are given explicitly in terms of $\{\mathbf{x}_{02}, R, R', R''\}$. The expectation of the queue length at an arbitrary time has the expansion

$$\mathbf{EX}(r) = r\mathbf{q}_1 + r^2\mathbf{q}_2 + o(r^2), \tag{14}$$

where

$$\mathbf{q}_1 = (\mathbf{x}'_{01} \otimes \theta, 0)R(I-R)^{-2},$$

$$\mathbf{q}_2 = (\mathbf{x}''_{01} \otimes \theta, 0)R(I-R)^{-2} + (\mathbf{x}'_{01} \otimes \theta, 0)[R(I-R)^{-2}R' + (I-R)^{-1}R'(I-R)^{-1}](I-R)^{-1}.$$

Proof. From (6) and Lemmas 3.1 and 3.2, $\lim_{r\to 0}\mathbf{X}(r)$ exists and $\mathbf{X}^{(n)}(r)$ is differentiable in r for all $n \geq 0$. For other details, see Appendix C. □

Lemma 3.1 and Theorems 3.3 and 3.4 imply that the basic quantities for the expansions of the stationary probability vector $X(r)$ of the Markov process $Q(r)$ are $\mathbf{x}_{02}$ and R. Other probabilities or descriptors can be expressed in terms of those two. But some resulting expressions turn out to be quite involved. Fortunately, by Theorems 3.3 and 3.4 those quantities can be expressed explicitly in those two terms. Therefore, we shall use not only $\{R, \mathbf{x}_{02}\}$ but also $\{R', R'', \mathbf{x}'_{01}, \mathbf{x}''_{01}, \mathbf{x}'_{02}, \mathbf{x}''_{02}\}$ in the next two sections.

4. Conditional queues. Theorem 3.4 shows that, when $r \to 0$, the mean queue length at an arbitrary time tends to 0. However, the queues observed at some special epochs are different. For example, the queue length at an arbitrary arrival or departure does not always tend to 0 when $r \to 0$. The conditional queue, such as the queue during a system-active period or during a busy period, may not be short when $r \to 0$. Furthermore, the queue observed by the last arrival of an input-active period is different from others. It is clear that the classical descriptors do not suffice to characterize these queues because some important information is carried by the conditional distributions of these queues. Therefore, more detailed discussion on those conditional queues is useful. We shall study the queue during a system-active period in detail. Others can be discussed similarly.

4.1. The queue during a system-active period. We denote by $\mathbf{Y}_a(r) = (\mathbf{y}_{a,0}(r), \mathbf{y}_{a,1}(r), \ldots)$ the probability vector of the distribution of the queue length in steady state, given that the system is active. Then we have

$$\mathbf{y}_{a,0}(r) = \frac{\mathbf{x}_{01}(r)}{1-\mathbf{x}_{02}(r)\mathbf{e}} \quad \text{and} \quad \mathbf{y}_{a,n}(r) = \frac{\mathbf{x}_n(r)}{1-\mathbf{x}_{02}(r)\mathbf{e}}, \quad n \geq 1.$$

Since $1-\mathbf{x}_{02}(r)\mathbf{e} = -r\mathbf{x}'_{02}\mathbf{e} - r^2\mathbf{x}''_{02}\mathbf{e} + o(r^2)$, where $\mathbf{x}'_{02}$ and $\mathbf{x}''_{02}$ are given in Appendix C, we have by Theorem 3.4 that

$$\mathbf{y}_{a,0}(r) = \frac{\mathbf{x}'_{01}}{(-\mathbf{x}'_{02}\mathbf{e})} + r\left[\frac{\mathbf{x}''_{01}}{(-\mathbf{x}'_{02}\mathbf{e})} + \frac{(\mathbf{x}''_{02}\mathbf{e})\mathbf{x}'_{01}}{(\mathbf{x}'_{02}\mathbf{e})^2}\right] + o(r),$$

$$\mathbf{y}_{a,n}(r) = \frac{(\mathbf{x}'_{01}\otimes\theta, 0)R^n}{(-\mathbf{x}'_{02}\mathbf{e})} + r\left[\frac{(\mathbf{x}''_{01}\otimes\theta, 0)R^n}{(-\mathbf{x}'_{02}\mathbf{e})} + \frac{(\mathbf{x}'_{01}\otimes\theta, 0)(\sum_{i=0}^{n-1} R^i R' R^{n-i-1})}{(-\mathbf{x}'_{02}\mathbf{e})} + \frac{(\mathbf{x}''_{02}\mathbf{e})(\mathbf{x}'_{01}\otimes\theta, 0)R^n}{(\mathbf{x}'_{02}\mathbf{e})^2}\right] + o(r), \quad n \geq 1. \tag{15}$$

We have two methods to obtain the limit distribution of $\mathbf{Y}_a(r)$ when $r \to 0$. The first one is to set $r = 0$ in (15). The second one is more fundamental and gives a probabilistic interpretation to $\mathbf{Y}_a(0)$; i.e., $\mathbf{Y}_a(0)$ is the stationary probability vector of a Markov process (see Theorem 4.2). To introduce the second method, we need the following lemma.

LEMMA 4.1.

(16)
$$\lim_{r\to 0} r(rD_{04} + D_{05})^{-1}D_{03} = \begin{cases} D_{04}^{-1}D_{03} & \text{if } D_{05} = 0, \\ -(\theta_4 D_{03}\mathbf{e})^{-1}\mathbf{e}\theta_4 D_{03} & \text{if } D_{05} \neq 0 \text{ and irreducible.} \end{cases}$$

Proof. See Appendix D. □

THEOREM 4.2. *The conditional probability distribution* $\{\mathbf{Y}_a(r)\}$ *converges, when* $r \to 0$, *to a positive probability vector* $\mathbf{Y}_a(0) = (\mathbf{y}_{a,0}, \ldots)$, *which is the stationary probability vector of the infinitesimal generator*

$$Q_a = \begin{pmatrix} D_{01} + D_{02}M & (D_{11}, D_{12}) \otimes \theta & & & \\ \begin{pmatrix} I \\ M \end{pmatrix} \otimes (C_1\mathbf{e}) & D_0 \oplus C_0 & D_1 \otimes I & & \\ & I \otimes C_1 & D_0 \oplus C_0 & D_1 \otimes I & \\ & & \ddots & \ddots & \ddots \end{pmatrix},$$

where $-M$ *is the limit matrix in* (16). $\mathbf{y}_{a,n} = \mathbf{y}_{a,0}(I \otimes \theta, 0)R^n$, $n \geq 1$, $\mathbf{y}_{a,0}$ $(= \mathbf{x}'_{01}/(-\mathbf{x}'_{02}\mathbf{e}))$ *satisfies*

$$\mathbf{y}_{a,0}[D_{01} + D_{02}M + (I \otimes \theta)(R_1 + R_2M) \otimes (C_1\mathbf{e})] = 0, \tag{17}$$

and $\mathbf{y}_{a,0}(I \otimes \theta, 0)(I - R)^{-1}\mathbf{e} = 1$.

Proof. From (6), we express $\mathbf{x}_{02}(r)$ in terms of $\mathbf{x}_{01}(r)$. Lemma 4.1 now leads to the result. □

THEOREM 4.3. *The expansion of the mean queue length given that the system is active is given by*

$$\mathbf{E}\mathbf{Y}_a(r) = \mathbf{q}_{a,0} + r\mathbf{q}_{a,1} + o(r), \tag{18}$$

where

$$\mathbf{q}_{a,0} = \frac{\mathbf{q}_1}{(-\mathbf{x}'_{02}\mathbf{e})},$$
$$\mathbf{q}_{a,1} = \frac{\mathbf{q}_2}{(-\mathbf{x}'_{02}\mathbf{e})} + \frac{(\mathbf{x}''_{02}\mathbf{e})\mathbf{q}_1}{(\mathbf{x}'_{02}\mathbf{e})^2}.$$

Proof. The mean queue length at an arbitrary time is

$$\mathbf{E}\mathbf{X}(r) = \mathbf{x}_0(r)(I \otimes \theta)R(r)(I - R(r))^{-2}\mathbf{e}.$$

The relation between the mean queue length at an arbitrary time and the mean queue length given that the system is active is given by

$$\mathbf{E}\mathbf{X}(r) = \mathbf{E}\mathbf{Y}_a(r)(1 - \mathbf{x}_{02}(r)). \tag{19}$$

We obtain (18) by Theorem 3.4 and simple calculations. □

$\mathbf{q}_{a,0}\mathbf{e}$ in Theorem 4.3 gives the expected number of customers in the system at an arbitrary time, given that the system is active, when r is very small. For bursty input, $\mathbf{q}_{a,0}\mathbf{e}$ can be very large (see Example 3 in section 7). Therefore, $\mathbf{q}_{a,0}$ provides more accurate information about the queue when the system is active than the average queue length at an arbitrary time, which tends to 0 when $r \to 0$.

4.2. Other conditional queue lengths. We denote by λ_e^* the arrival rate of the last arrivals of input-active periods. Then $\lambda_e^*(r) = \theta_{11}(r)D_{12}\mathbf{e}$. We denote by $\mathbf{X}_e(r) = (\mathbf{x}_{e,0}(r), \mathbf{x}_{e,1}(r), \ldots)$ the conditional probability vector of the queue length at the arrival epoch of the last arrival of an arbitrary input-active period. Then we have

$$\mathbf{x}_{e,0}(r) = \frac{1}{\lambda_e^*(r)}\mathbf{x}_{01}(r)(0, D_{12}),$$

$$\mathbf{x}_{e,n}(r) = \frac{1}{\lambda_e^*(r)}\mathbf{x}_n(r)(\bar{D}_{12} \otimes I), \quad n \geq 1,$$

where $\bar{D}_{12} = \left(\begin{smallmatrix} 0 & D_{12} \\ 0 & 0 \end{smallmatrix}\right)$. Similar to $\lambda^*(r)$, $\lambda_e^*(r)$ has the expansion

$$\lambda_e^*(r) = -r\theta_{12}D_{03}(D_{01} + D_{11})^{-1}D_{12}\mathbf{e} - r^2\theta'_{12}D_{03}(D_{01} + D_{11})^{-1}D_{12}\mathbf{e} + o(r^2),$$

where θ_{12} and θ'_{12} are given in Appendix B. By routine calculations, we obtain the following expansion:

$$\mathbf{E}\mathbf{X}_e(r) = \mathbf{q}_{e,0} + r\mathbf{q}_{e,1} + o(r), \tag{20}$$

where

$$\mathbf{q}_{e,0} = \frac{-1}{(\theta_{12}D_{03}(D_{01} + D_{11})^{-1}D_{12}\mathbf{e})}\mathbf{q}_1(\bar{D}_{12} \otimes I),$$

$$\mathbf{q}_{e,1} = \frac{-1}{(\theta_{12}D_{03}(D_{01} + D_{11})^{-1}D_{12}\mathbf{e})}\left[\mathbf{q}_2 - \frac{\theta'_{12}D_{03}(D_{01} + D_{11})^{-1}D_{12}\mathbf{e}}{(\theta_{12}D_{03}(D_{01} + D_{11})^{-1}D_{12}\mathbf{e})}\mathbf{q}_1\right](\bar{D}_{12} \otimes I).$$

For other conditional queues, we only give their relations with $\mathbf{X}(r)$. The corresponding expansions can be obtained similarly to Theorem 4.3 or (20). We omit the details, but state the following results.

Let $\mathbf{Y}_b(r) = (\mathbf{y}_{b,1}(r), \mathbf{y}_{b,2}(r), \ldots)$ be the conditional probability vector of the queue length, given that the server is busy. Then we have

$$\mathbf{y}_{b,n}(r) = \frac{\mathbf{x}_n(r)}{1 - \mathbf{x}_0(r)\mathbf{e}}, \quad n \geq 1. \tag{21}$$

Let $\mathbf{X}_a(r) = (\mathbf{x}_{a,0}(r), \mathbf{x}_{a,1}(r), \ldots)$ be the conditional probability vector of the queue length at an arbitrary arrival. Then

$$\mathbf{x}_{a,0}(r) = \frac{1}{\lambda^*(r)}\mathbf{x}_{01}(r)(D_{11}, D_{12}),$$

(22)

$$\mathbf{x}_{a,n}(r) = \frac{1}{\lambda^*(r)}\mathbf{x}_n(r)(D_1 \otimes I), \quad n \geq 1.$$

Let $\mathbf{X}_d(r) = (\mathbf{x}_{d,0}(r), \mathbf{x}_{d,1}(r), \ldots)$ be the conditional probability vector of the queue length immediately after a departure. Then

$$\mathbf{x}_{d,n}(r) = \frac{1}{\lambda^*(r)}\mathbf{x}_{n+1}(r)(I \otimes C_1), \quad n \geq 0. \tag{23}$$

5. Waiting times. As the queue length at an arbitrary arrival, the virtual waiting time $w(r)$ is not adequate to characterize the waiting process in an episodic queue. We are interested in conditional waiting times such as the waiting time during a system-active period $w_a(r)$, the waiting time during a busy period, the waiting time of the last arrival of an input-active period, etc. For brevity, we give details only for the virtual waiting time and the virtual waiting time during a system-active period. Other conditional waiting times can be similarly obtained.

From (13) and (15), we obtain that

$$\mathbf{E}e^{-sw(r)} = 1 + r\mathbf{x}_0'(I\otimes\theta)f^*(s)\mathbf{e} + r^2[\mathbf{x}_0''(I\otimes\theta)f^*(s)\mathbf{e} + \mathbf{x}_0'(I\otimes\theta)f_1^*(s)]\mathbf{e} + o(r^2), \tag{24}$$

$$\mathbf{E}e^{-sw_a(r)} = \frac{\mathbf{x}_0'(I\otimes\theta)f^*(s)\mathbf{e} - \mathbf{x}_{02}'\mathbf{e}}{(-\mathbf{x}_{02}\mathbf{e})} + r\left[\left(\frac{(\mathbf{x}_{02}''\mathbf{e})\mathbf{x}_0'}{(\mathbf{x}_{02}'\mathbf{e})^2} + \frac{\mathbf{x}_0''}{(-\mathbf{x}_{02}'\mathbf{e})}\right)\cdot(I\otimes\theta)f^*(s) + \frac{\mathbf{x}_0'(I\otimes\theta)f_1^*(s)}{(-\mathbf{x}_{02}'\mathbf{e})}\right]\mathbf{e} + o(r), \tag{25}$$

where

$$f^*(s) = \sum_{n=0}^{\infty} R^n[h^*(s)]^n,$$
$$f_1^*(s) = \sum_{n=1}^{\infty}\left(\sum_{i=0}^{n-1} R^i R' R^{n-1-i}\right)[h^*(s)]^n,$$

and $h^*(s) = (sI - D\oplus C_0)^{-1}(I\otimes C_1)$.

LEMMA 5.1. *The matrices $f^*(0)$ and $f_1^*(0)$ are obtained by solving the equation*

$$X = X_0 + RX(-D\oplus C_0)^{-1}(I\otimes C_1) \tag{26}$$

with $X_0 = I$ and $X_0 = R'f^*(0)(-D\oplus C_0)^{-1}(I\otimes C_1)$, respectively. A general solution for (26) *is given by*

$$\phi(X) = \phi(X_0)(I - R^T\otimes[(-D\oplus C_0)^{-1}(I\otimes C_1)])^{-1}. \tag{27}$$

In addition, we have that

$$f^{*\prime}(0)\mathbf{e} = (I-R)^{-1}Rf^*(0)(D\oplus C_0)^{-1}\mathbf{e},$$

$$f_1^{*\prime}(0)\mathbf{e} = (I-R)^{-1}[Rf_1^*(0) + R'(I-R)^{-1}f^*(0)](D\oplus C_0)^{-1}\mathbf{e}.$$

Proof. By the definitions of $f^*(s)$ and $f_1^*(s)$, we have that

$$f^*(s) = I + Rf^*(s)h^*(s),$$
$$f_1^*(s) = R'f^*(s)h^*(s) + Rf_1^*(s)h^*(s),$$

and simple calculations lead to the stated results. □

THEOREM 5.2.

$$\mathbf{E}w(r) = rw_1 + r^2 w_2 + o(r^2), \tag{28}$$

$$\mathbf{E}w_a(r) = w_{a,0} + rw_{a,1} + o(r), \tag{29}$$

where

$$\begin{aligned} w_1 &= (\mathbf{x}'_{01} \otimes \theta, 0)R(I-R)^{-1} f^*(0)(-D \oplus C_0)^{-1}\mathbf{e}, \\ w_2 &= (\mathbf{x}''_{01} \otimes \theta, 0)R(I-R)^{-1} f^*(0)(-D \oplus C_0)^{-1}\mathbf{e} + (\mathbf{x}'_{01} \otimes \theta, 0) \\ &\quad \cdot (I-R)^{-1}[Rf_1^*(0) + R'(I-R)^{-1} f^*(0)](-D \oplus C_0)^{-1}\mathbf{e}, \\ w_{a,0} &= \frac{w_1}{(-\mathbf{x}'_{02}\mathbf{e})}, \qquad w_{a,1} = \frac{w_2}{(-\mathbf{x}'_{02}\mathbf{e})} + \frac{(\mathbf{x}''_{02}\mathbf{e})w_1}{(\mathbf{x}'_{02}\mathbf{e})^2}. \end{aligned}$$

w_1 and $w_{a,0}$ are nonnegative.

Proof. Differentiating both sides of (24) and (25) with respect to s, by Lemma 5.1, we obtain (28) and (29). Since $\mathbf{x}'_{01} \geq 0$ and $-\mathbf{x}'_{02}\mathbf{e} > 0$, w_1 and $w_{a,0}$ are nonnegative. □

6. The busy period and active periods. The busy period is an important descriptor for queueing systems. We refer to Hsu and He [3], Neuts [6, 8], and Ramaswami [10] for studies on the busy period of queues related to quasi-birth-and-death processes. For episodic queues, we are also interested in the system-active period and the input-active period. The relations among the busy period, the system-active period, and the input-active period contain much information about the episodic queue.

6.1. The busy period. Let $N_b(r)$ be the number of customers served in a busy period and $\tau_b(r)$ be the length of the busy period. We define

$$\begin{aligned} g_{(i_0,j_0)(i,j)}(k,x,r) = \mathbf{P}\{&N_b(r) = k, \tau_b(r) \leq x, J_1(\tau_b(r)) = i, \\ &J_2(\tau_b(r)) = j | J_1(0) = i_0, J_2(0) = j_0\} \end{aligned}$$

for $k \geq 0,\ x > 0$. Let $G^*(z,s,r)$ be the joint transform of the matrix with elements $g_{(i_0,j_0)(i,j)}(k,x,r)$. By Theorem 3.3.1 in Neuts [6], $G^*(z,s,r)$ is the unique solution to the equation

$$G^*(z,s,r) = [sI - D_0(r) \oplus C_0]^{-1}\{z(I \otimes C_1) + (D_1 \otimes I)[G^*(z,s,r)]^2\} \tag{30}$$

for $0 \leq z < 1,\ s \geq 0$. Let $G(r) = \lim_{z\to 1, s\to 0} G^*(z,s,r)$. $G(r)$ is the minimal nonnegative solution to (30) with $s = 0$ and $z = 1$. Since $\rho(r) < 1$, $G(r)\mathbf{e} = \mathbf{e}$. Furthermore, it is clear that $G = \lim_{r\to 0} G(r)$ exists. G is the minimal nonnegative solution to equation (30) with $s = 0$, $z = 1$, and $r = 0$ and has the structure

$$G = \begin{pmatrix} G_1 & G_2 \\ 0 & G_3 \end{pmatrix}$$

with G_1, G_2, and G_3 satisfying

$$0 = I \otimes C_1 + (D_{01} \oplus C_0)G_1 + (D_{11} \otimes I)G_1^2, \tag{31}$$

$$\begin{aligned} 0 = {} & (D_{02} \otimes I)G_3 + [D_{01} \oplus C_0 + (D_{11} \otimes I)G_1]G_2 \\ & + (D_{11} \otimes I)G_2G_3 + (D_{12} \otimes I)G_3^2, \end{aligned} \tag{32}$$

$$0 = I \otimes C_1 + (D_{05} \oplus C_0)G_3. \tag{33}$$

G_1 is the minimal nonnegative solution to (31) and $G_3 = -(D_{05} \oplus C_0)^{-1}(I \otimes C_1)$. Let $V_{11} = -[D_{01} \oplus C_0 + (D_{11} \otimes I)G_1]$. Since $-V_{11}\mathbf{e} < 0$ and $D(r)$ and $Q(r)$ are irreducible, V_{11} is invertible. Let γ_1 be the eigenvector corresponding to the eigenvalue with the largest real part of $D_{01} + D_{11}$ and let $\gamma = \gamma_1 \otimes \theta_2$. Then we have that

$$0 < \gamma[(D_{01} \oplus C_0)(G_1 - I) + (D_{11} \otimes I)(G_1^2 - I)]$$

or, equivalently, since $I - G_1$ is invertible,

$$0 > \gamma[D_{01} \oplus C_0 + (D_{11} \otimes I)(G_1 + I)].$$

So, $\gamma > \gamma(D_{11} \otimes I)V_{11}^{-1}$, which implies $\mathrm{sp}((D_{11} \otimes I)V_{11}^{-1}) < 1$. Since $G_3\mathbf{e} = \mathbf{e}$, we can apply the ϕ-transform to equation (32) to derive G_2.

LEMMA 6.1. *The derivative $G^{(n)}(r)$ is differentiable in r for $n \geq 0$. $G(r)$ has the following expansion at $r = 0$:*

$$G(r) = G + rG' + r^2G'' + o(r^2), \tag{34}$$

where G' and G'' satisfy the equations

$$\begin{aligned} 0 &= (\bar{D}_0 \otimes I)G - VG' + (D_1 \otimes I)G'G, \\ 0 &= (\bar{D}_0 \otimes I)G' + (D_1 \otimes I)(G')^2 - VG'' + (D_1 \otimes I)G''G \end{aligned} \tag{35}$$

with $V = -(D_0 \oplus C_0 + (D_1 \otimes I)G)$. G' and G'' can be solved explicitly by applying the ϕ-transform to (35).

Proof. The differentiability of $G^{(n)}(r)$ follows by Theorem 3.2 in He [2]. Equations (35) follow from (30). Since $-V\mathbf{e} = (I \otimes C_0)\mathbf{e} < 0$ and $D(r)$ and C are irreducible, V is invertible. By routine calculations, we have

$$-V + D_1 \otimes I = [D \oplus C + (D_1 \otimes I - I \otimes C_1)\mathbf{eg}](I - G + \mathbf{eg})^{-1},$$

where $\mathbf{g}$ is the left invariant vector of the matrix G. So,

$$(\theta_1(0) \otimes \theta_2)(-V + D_1 \otimes I) = (\lambda^* - \mu^*)\mathbf{g} < 0.$$

Then $\theta_1(0) \otimes \theta_2 > (\theta_1(0) \otimes \theta_2)(D_1 \otimes I)V^{-1}$ implies that the Perron–Frobenius eigenvalue of $(D_1 \otimes I)V^{-1}$ is less then 1. Hence, the ϕ-transform method can be applied to solve equations in (35) for G' and G''. ∎

Let $\mathbf{u}(r)$ and $\mathbf{v}(r)$ denote, respectively, the mean number of customers served during a busy period and the mean length of the busy period. Then

$$\mathbf{u}(r) = \frac{\partial G^*(z,s,r)\mathbf{e}}{\partial z}\Big|_{z=1,s=0} \quad \text{and} \quad \mathbf{v}(r) = -\frac{\partial G^*(z,s,r)\mathbf{e}}{\partial s}\Big|_{z=1,s=0}.$$

From Neuts [6], $\mathbf{u}(r)$ and $\mathbf{v}(r)$ are explicitly given by

$$\begin{aligned} \mathbf{u}(r) &= -[D(r) \oplus C_0 + (D_1 \otimes I)G(r)]^{-1}(I \otimes C_1)\mathbf{e}, \\ \mathbf{v}(r) &= -[D(r) \oplus C_0 + (D_1 \otimes I)G(r)]^{-1}\mathbf{e}. \end{aligned} \tag{36}$$

THEOREM 6.2. *The vector* $\mathbf{u}(r)$ *has the expansion*

$$\mathbf{u}(r) = \mathbf{u}(0) + r\mathbf{u}'(0) + r^2\mathbf{u}''(0) + o(r^2), \tag{37}$$

where $\mathbf{u}(0)$ *is obtained from* (36) *by setting* $r = 0$ *and*

$$\begin{aligned}\mathbf{u}'(0) &= -[D \oplus C_0 + (D_1 \otimes I)G]^{-1}[\bar{D}_0 \otimes I + (D_1 \otimes I)G']\mathbf{u}(0),\\ \mathbf{u}''(0) &= -[D \oplus C_0 + (D_1 \otimes I)G]^{-1}[(D_1 \otimes I)G''\mathbf{u}(0)\\ &\qquad +[\bar{D}_0 \otimes I + (D_1 \otimes I)G']\mathbf{u}'(0)].\end{aligned}$$

$\mathbf{v}(r)$ *has the expansion* $\mathbf{v}(r) = \mathbf{v}(0) + r\mathbf{v}'(0) + r^2\mathbf{v}''(0) + o(r^2)$, *where* $\mathbf{v}(0)$ *is obtained from* (36) *by setting* $r = 0$, $\mathbf{v}'(0)$ *and* $\mathbf{v}''(0)$ *are obtained from the above formulas by replacing* $\mathbf{u}(0)$ *by* $\mathbf{v}(0)$, $\mathbf{u}'(0)$ *by* $\mathbf{v}'(0)$, *and* $\mathbf{u}''(0)$ *by* $\mathbf{v}''(0)$, *accordingly.*

Proof. Replacing $G(r)$ by its expansion, we have

$$\begin{aligned}&[D(r) \oplus C_0 + (D_1 \otimes I)G(r)]^{-1}\\ &\quad = \{D \oplus C_0 + (D_1 \otimes I)G + r[\bar{D}_0 \otimes I + (D_1 \otimes I)G']\\ &\qquad +r^2(D_1 \otimes I)G'' + o(r^2)\}^{-1}.\end{aligned}$$

The stated formula follows by direct calculations. □

6.2. The number of input-active periods during a busy period. To study the relation among the busy and input-active periods, it is natural to ask how many input-active periods there are in a busy period (strictly within a busy period). We denote this random variable by $\eta_{b,a}(r)$. We define

$$\begin{aligned}g_{(i_0,j_0)(i,j)}(k_1,k_2,x,r) = \mathbf{P}\{&\tau_b(r) \le x, N_b(r) = k_1, \eta_{b,a}(r) = k_2, J_1(\tau_b(r)) = i,\\ &J_2(\tau_b(r)) = j | J_1(0) = i_0, J_2(0) = j_0\}\end{aligned}$$

for $k_1, k_2 \ge 0$, $x \ge 0$, and $r \ge 0$.

THEOREM 6.3. *Let* $G^*(z_1, z_2, s, r)$ *be the joint transform of the matrix with elements* $g_{(i_0,j_0)(i,j)}(k_1,k_2,x,r)$. $G^*(z_1, z_2, s, r)$ *is the unique solution to the following equation:*

$$\begin{aligned}G^*(z_1,z_2,s,r) = [sI - (\tilde{D}_0 + r\bar{D}_0 + z_2\bar{D}_{02}) \oplus C_0]^{-1}\{&z_1(I \otimes C_1)\\ &+((\bar{D}_{11} + z_2\bar{D}_{12}) \otimes I)[G^*(z_1,z_2,s,r)]^2\},\end{aligned} \tag{38}$$

where

$$\tilde{D}_0 = \begin{pmatrix} D_{01} & 0 \\ 0 & D_{05} \end{pmatrix},\ \bar{D}_{02} = \begin{pmatrix} 0 & D_{02} \\ 0 & 0 \end{pmatrix},\ \bar{D}_{11} = \begin{pmatrix} D_{11} & 0 \\ 0 & 0 \end{pmatrix}.$$

Proof. The basic idea of the proof is similar to that of Lemmas 3.3.1 and 3.3.2 and Theorem 3.3.1 in Neuts [6]. Here we give a brief probabilistic discussion only.

Notice that a transition with matrix $I \otimes C_1$ ends a service, a transition with matrix $\bar{D}_{02} \otimes I$ terminates an input-active period, a transition with matrix $\bar{D}_{11} \otimes I$ brings a new customer, while a transition with matrix $\bar{D}_{12} \otimes I$ brings a new customer and ends an input-active period. Let $G(z_1, z_2, x, r)$ be the matrix whose elements are the joint generating functions of $g_{(i_0,j_0)(i,j)}(k_1,k_2,x,r)$ with respect to k_1 and

k_2. By conditioning on the first transition, which either ends a service or an input-active period, or brings a new customer, or brings a new customer and terminates an input-active period, we have that

$$\begin{aligned} G(z_1,z_2,x,r) &= \int_0^x \exp\{t(\tilde{D}_0 + r\bar{D}_0) \oplus C_0\} z_1 (I \otimes C_1) dt \\ &\quad + \int_0^x \exp\{t(\tilde{D}_0 + r\bar{D}_0) \oplus C_0\} z_2 (\bar{D}_{02} \otimes I) G(z_1,z_2,x-t,r) dt \\ &\quad + \int_0^x \exp\{t(\tilde{D}_0 + r\bar{D}_0) \oplus C_0\} ((\bar{D}_{11} + z_2 \bar{D}_{12}) \otimes I) \\ &\qquad \cdot \int_0^{x-t} G(z_1,z_2,du,r) G(z_1,z_2,x-t-u,r) dt. \end{aligned}$$

Taking the Laplace–Stieltjes (L.S.) transform with respect to x of that equation we obtain (38). □

Set $G_a^*(z,r) = G^*(1,z,0,r)$. $G_a^*(z,r)$ is the generating function of the number of input-active periods in a busy period and satisfies (38) with $z_1 = 1$ and $s = 0$. Let

$$\mathbf{u}_{b,a}(r) = \left[\frac{\partial G_a^*(z,r)}{\partial z}|_{z=1} \right] \mathbf{e}.$$

For $r = 0$, it is easy to show that

$$G_a^*(z,0) = \begin{pmatrix} G_1 & zG_2 \\ 0 & G_3 \end{pmatrix} \quad \text{and} \quad \mathbf{u}_{b,a}(0) = \begin{pmatrix} G_2\mathbf{e} \\ 0 \end{pmatrix}. \tag{39}$$

Therefore, when $r \to 0$, the expected number of input-active periods in a busy period tends to 1.

THEOREM 6.4. *The mean number of input-active periods in a busy period is given by*

$$\mathbf{u}_{b,a}(r) = -[D(r) \oplus C_0 + (D_1 \otimes I)G(r)]^{-1}((\bar{D}_{02} + \bar{D}_{12}) \otimes I)\mathbf{e}, \tag{40}$$

and $\mathbf{u}_{b,a}(r)$ *has the expansion*

$$\mathbf{u}_{b,a}(r) = \mathbf{u}_{b,a}(0) + r\mathbf{u}'_{b,a}(0) + r^2\mathbf{u}''_{b,a}(0) + o(r^2), \tag{41}$$

where

$$\begin{aligned} \mathbf{u}'_{b,a}(0) &= -[D \oplus C_0 + (D_1 \otimes I)G]^{-1}[(D_1 \otimes I)G' + \bar{D}_0 \otimes I]\mathbf{u}_{b,a}(0), \\ \mathbf{u}''_{b,a}(0) &= -[D \oplus C_0 + (D_1 \otimes I)G]^{-1}[(D_1 \otimes I)G''\mathbf{u}_{b,a}(0) \\ &\qquad + [(D_1 \otimes I)G' + \bar{D}_0 \otimes I]\mathbf{u}'_{b,a}(0)]. \end{aligned} \tag{42}$$

Proof. It is similar to Theorem 6.2. □

Note 2. By the same method, we can derive similar formulas for the total number of input-active periods in, or overlapping with, a busy period.

6.3. The number of busy periods in an input-active period. The number of busy periods during an input-active period shows how often the server gets idle when the arrival process is active. To study that random variable, we consider the absorbing Markov process

$$\bar{Q} = \begin{pmatrix} 0 & 0 & 0 & & & \\ (D_{02}+D_{12})\mathbf{e} & D_{01} & D_{11}\otimes\theta & & & \\ (D_{02}+D_{12})\mathbf{e}\otimes\mathbf{e} & I\otimes C_1\mathbf{e} & D_{01}\oplus C_0 & D_{11}\otimes I & & \\ (D_{02}+D_{12})\mathbf{e}\otimes\mathbf{e} & & I\otimes C_1 & D_{01}\oplus C_0 & D_{11}\otimes I & \\ \vdots & & & \ddots & \ddots & \ddots \end{pmatrix},$$

whose first state (or level) is absorbing. The time until absorption into the first state of the Markov process $\bar{Q}$ is the duration of an input-active period of the queue. The number of transitions from level 2 to level 1 prior to absorption in the Markov process $\bar{Q}$ is the number of busy periods during an input-active period of the queue.

Let $\tilde{G}_1^*(z,s)$ be the joint transform of the first passage time starting from level i $(i>2)$ to level $i-1$ and the number of customers served during that time. Similarly, let $\tilde{G}_2^*(z,s)$ be the joint transform of the first passage time starting from level i (>1) to level 0 without entering level $i-1$ and the number of customers served during that time. We have equations:

$$\tilde{G}_1^*(z,s) = (sI - D_{01}\oplus C_0)^{-1}\{z(I\otimes C_1) + (D_{11}\otimes I)[\tilde{G}_1^*(z,s)]^2\}, \tag{43}$$

$$\tilde{G}_2^*(z,s) = (sI - D_{01}\oplus C_0)^{-1}\{z((D_{02}+D_{12})\mathbf{e})\otimes\mathbf{e} + (D_{11}\otimes I) \cdot[I + \tilde{G}_1^*(z,s)]\tilde{G}_2^*(z,s)\}. \tag{44}$$

We denote by $\tilde{G}_i = \lim_{z\to 1, s\to 0}\tilde{G}_i^*(z,s)$, $i=1,2$; then $\tilde{G}_1 = G_1$ and

$$\tilde{G}_2 = -[(D_{01}+D_{11})\oplus C_0 + (D_{11}\otimes I)G_1]^{-1}[((D_{02}+D_{12})\mathbf{e})\otimes\mathbf{e}].$$

LEMMA 6.5. *The Perron–Frobenius eigenvalue of the matrix* $-[D_{01}\oplus C_0 + (D_{11}\otimes I)\tilde{G}_1]^{-1}(D_{11}\otimes I)$ *is less than* 1 *and* $\tilde{G}_2 + G_1\mathbf{e} = \mathbf{e}$.

Proof. The first result is obvious. Postmultiplying by $\mathbf{e}$ on both sides of equation (43) and adding them to equation (44) yields

$$\tilde{G}_1\mathbf{e} + \tilde{G}_2 - \mathbf{e} = -[D_{01}\oplus C_0 + (D_{11}\otimes I)\tilde{G}_1]^{-1}(D_{11}\otimes I)(\tilde{G}_1\mathbf{e} + \tilde{G}_2 - \mathbf{e}).$$

By the first result, $\tilde{G}_1\mathbf{e} + \tilde{G}_2 - \mathbf{e} = 0$. □

We denote by τ_a the length of an input-active period, by N_{τ_a} the number of customers served during the input-active period, and by $\eta_{a,b}$ the number of busy periods during the input-active period. We assume that there is no customer in the system at the beginning of the input-active period. Let $f_i^*(z,s,n) = \mathbf{E}\{e^{-s\tau_a}z^{N_{\tau_a}}I_{\{\eta_{a,b}=n\}}|J_1(0) = i\}, n\geq 0$, and $\mathbf{f}^*(z,s,n) = (f_1^*(z,s,n),\ldots,f_{m_1}^*(z,s,n))^T$, where I_A is the indicator of the event A. We have that

$$\mathbf{f}^*(z,s,n) = [(sI - D_{01})^{-1}(D_{11}\otimes\theta)\tilde{G}_1^*(z,s)(I\otimes\mathbf{e})]^n(sI - D_{01})^{-1} \cdot[(D_{02}+D_{12})\mathbf{e} + (D_{11}\otimes\theta)\tilde{G}_2^*(z,s)], \quad n\geq 0. \tag{45}$$

THEOREM 6.6. *We assume that there is no customer in the queueing system at the beginning of an input-active period. The joint density of the arrival phase and the number of busy periods within the input-active period is given by*

$$\mathbf{p}_n = -[-D_{01}^{-1}(D_{11}\otimes\theta)G_1(I\otimes\mathbf{e})]^n D_{01}^{-1} \cdot[(D_{02}+D_{12})\mathbf{e}+D_{11}\mathbf{e}-(D_{11}\otimes\theta)G_1\mathbf{e}], \quad n\geq 0. \tag{46}$$

The vector $\mathbf{u}_{a,b}=\sum_{n=1}^{\infty} n\mathbf{p}_n$ *is given by*

$$\mathbf{u}_{a,b} = -D_{01}^{-1}(D_{11}\otimes\theta)G_1(I\otimes\mathbf{e})[I+D_{01}^{-1}(D_{11}\otimes\theta)G_1(I\otimes\mathbf{e})]^{-2} \cdot D_{01}^{-1}[(D_{02}+D_{12})\mathbf{e}+D_{11}\mathbf{e}-(D_{11}\otimes\theta)G_1\mathbf{e}]. \tag{47}$$

Proof. These results are obtained by routine calculations. The matrix $I+G_1(I\otimes\mathbf{e})D_{01}^{-1}(D_{11}\otimes\theta)$ is invertible since $-[G_1(I\otimes\mathbf{e})D_{01}^{-1}(D_{11}\otimes\theta)]\mathbf{e}<G_1\mathbf{e}<\mathbf{e}$. By Lemma 6.5, it is easily verified that $\{\alpha\mathbf{p}_n, n\geq 0\}$ is a proper probability density for any probability vector α. □

Note 3. From Theorem 6.6, it is not difficult to derive the distribution of the number of busy periods in an input-active period for the case when there are customers in the system at the beginning of the input-active period. For example, the mean number of busy periods during an input-active period, given that there are k (>0) customers in the system initially, is given by

$$\mathbf{u}_{a,b} = -G_1^{k-1}[I+G_1(I\otimes\mathbf{e})D_{01}^{-1}(D_{11}\otimes\theta)]^{-2}G_1(I\otimes\mathbf{e})D_{01}^{-1} \cdot[(D_{02}+D_{11}+D_{12})\mathbf{e}-(D_{11}\otimes\theta)G_1]. \tag{48}$$

6.4. The system-active period. We now give a brief discussion of the system-active period.

Let $\bar{G}_1^*(z_1,z_2,z_3,s,r)$ be the joint transform of the number of customers served, the number of input-active periods, the number of busy periods during a first passage to sublevel 0_2, and the length of that first passage, given that the process starts from sublevel 0_1. Similarly, we define $\bar{G}_2^*(z_1,z_2,z_3,s,r)$ for transitions from level 1 to sublevel 0_2. Then we have the following basic equations (for brevity, we omit (z_1,z_2,z_3,s,r)):

$$\bar{G}_1^* = (sI-D_{01})^{-1}[z_2D_{02}+((D_{11},z_2D_{12})\otimes\theta)\bar{G}_2^*], \tag{49}$$

$$\bar{G}_2^* = [sI-(\tilde{D}_0+r\bar{D}_0+z_2\bar{D}_{02})\oplus C_0]^{-1}\left\{z_1z_3\begin{pmatrix}0\\ I\otimes(C_1\mathbf{e})\end{pmatrix} +z_1z_3\begin{pmatrix}I\otimes(C_1\mathbf{e})\\ 0\end{pmatrix}\bar{G}_1^*+((\bar{D}_{11}+z_2\bar{D}_{12})\otimes I)G^*\bar{G}_2^*\right\}, \tag{50}$$

where G^* is the abbreviation for the matrix $G^*(z_1,z_2,s,r)$.

Analogous to sections 6.1 and 6.2, some explicit formulas can be derived for the mean number of customers served, mean number of input-active periods, and mean number of busy periods in a system-active period as well as their corresponding expansions. For brevity, we omit the details.

7. Numerical results. In this section, we discuss numerical results for two examples. We pay special attention to the relation between the burstiness of the arrivals and services and the derivatives of the vector $\mathbf{x}_0(r)$ at $r=0$ and the coefficients of the

various expansions. We are also concerned about the efficiency and accuracy of the approximation. To measure the accuracy of the approximation, we use the following formula:

$$\mathcal{E} = \frac{|Real - Approx|}{Real}. \tag{51}$$

Figures presented in this section show the graphs of $\mathcal{E}$ as a function of r for various descriptors.

Example 3. In this example, the input process has coefficient matrices

$$D_{01} = \begin{pmatrix} -50. & 1. \\ 1. & -100. \end{pmatrix}, \quad D_{02} = \begin{pmatrix} 1. & 0. \\ 0. & 1. \end{pmatrix},$$
$$D_{03} = \begin{pmatrix} 0.2 & 0.05 \\ 0.5 & 0.5 \end{pmatrix}, \quad D_{04} = \begin{pmatrix} -0.5 & 0.25 \\ 0.0 & -1. \end{pmatrix}, \quad D_{05} = \begin{pmatrix} -5. & 5. \\ 2. & -2. \end{pmatrix},$$
$$D_{11} = \begin{pmatrix} 48. & 0. \\ 0. & 98. \end{pmatrix}, \quad D_{12} = \begin{pmatrix} 0. & 0. \\ 0. & 0. \end{pmatrix}.$$

This is a bursty stochastic process. The service process has five choices: a bursty *MAP* with coefficient matrices

$$C_0 = \begin{pmatrix} -20. & 1. \\ 0.5 & -1. \end{pmatrix}, \quad C_1 = \begin{pmatrix} 18. & 1. \\ 0.25 & 0.25 \end{pmatrix};$$

an *MMPP* with coefficient matrices

$$C_0 = \begin{pmatrix} -10. & 3. \\ 1. & -3. \end{pmatrix}, \quad C_1 = \begin{pmatrix} 7. & 0. \\ 0. & 2. \end{pmatrix};$$

a *PH*-renewal process with coefficient matrices

$$C_0 = \begin{pmatrix} -5. & 3. \\ 1. & -2. \end{pmatrix}, \quad C_1 = \begin{pmatrix} 0.6 & 1.4 \\ 0.3 & 0.7 \end{pmatrix};$$

a Poisson process with $\mu^* = 1$; and an Erlang process with $k = 5$ and $\mu^* = 0.2$. According to the simulation of the five service processes, their subjective order of burstiness is *MAP*, *MMPP*, *PH*, *Poisson*, and *Erlang*, respectively.

Table 1 lists the numerical values of some coefficients in the expansions of the expected queue lengths and waiting times, at an arbitrary time and during a system-active period, respectively. Most of these coefficients are large in absolute value, which is obviously a result of the burstiness of the input process. So, the variation of the queueing system caused by a small change in r is large. For example, a large value of $\mathbf{q}_1\mathbf{e}$ means that the expected number of customers in the queue has a high growth rate as a function of r (when r is small). While the expected queue length at an arbitrary time is nearly 0 for small r, $\mathbf{q}_{a0}\mathbf{e} \approx 76$ approximately gives the expected number of customers in the queue when the system is active. This difference shows the importance of the system-active period for an episodic queue.

TABLE 1
Expansion coefficients of Example 3.

	Erlang	*Poisson*	*PH*	*MMPP*	*MAP*
$\mathbf{x}'_{01}\mathbf{e}$	0.0124	0.0124	0.0124	0.0124	0.0123
$\mathbf{x}''_{01}\mathbf{e}$	−0.70	−0.70	−0.71	−0.71	−0.75
$\mathbf{x}'_{02}\mathbf{e}$	−57.01	−57.01	−57.01	−57.15	−61.28
$\mathbf{x}''_{02}\mathbf{e}$	47.55	47.55	47.55	55.16	308.10
$\mathbf{q}_1\mathbf{e}$	4316.63	4316.63	4316.63	4326.73	4630.95
$\mathbf{q}_2\mathbf{e}$	239174.07	239174.07	239215.09	240344.69	253223.97
$\mathbf{q}_{a0}\mathbf{e}$	75.71	75.71	75.71	75.71	75.57
$\mathbf{q}_{a1}\mathbf{e}$	4258.27	4258.27	4258.99	4278.62	4512.45
w_1	4316.63	4316.63	4317.34	4347.16	4904.44
w_2	239174.04	239174.04	239214.45	240336.00	251434.47
w_{a0}	75.71	75.71	75.72	76.07	80.04
w_{a1}	4258.27	4258.27	4258.99	4278.81	4505.69

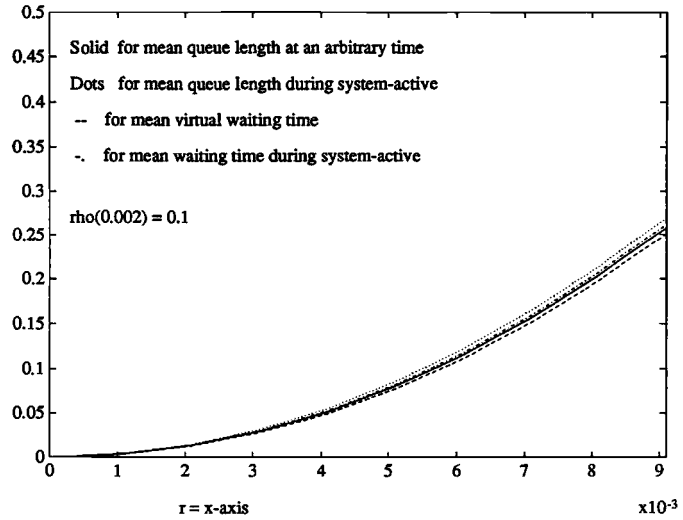


FIG. 1. *Example* 3. *Difference curves for the MAP service.*

Figures 1 and 2 show the relative differences between the approximate and the exact values of some quantities discussed in sections 2, 3, 4, and 5. They show that the approximation is quite good for traffic intensities $\rho(r)$ up to 0.1.

Discussions of the approximations of the busy and active periods-related quantities are similar to those of queue lengths and waiting times. Our numerical experiments show that the number of customers served in a busy period is very large since the input process is bursty. The number of the input-active periods in a busy period is close to 1 as is the number of busy periods during a system-active period. In fact, this last number is close to 1 for traffic intensity up to 1. An interesting observation is that the mean length of the input-active period is much smaller than the mean duration of the busy period. The input-active period typically ends with a

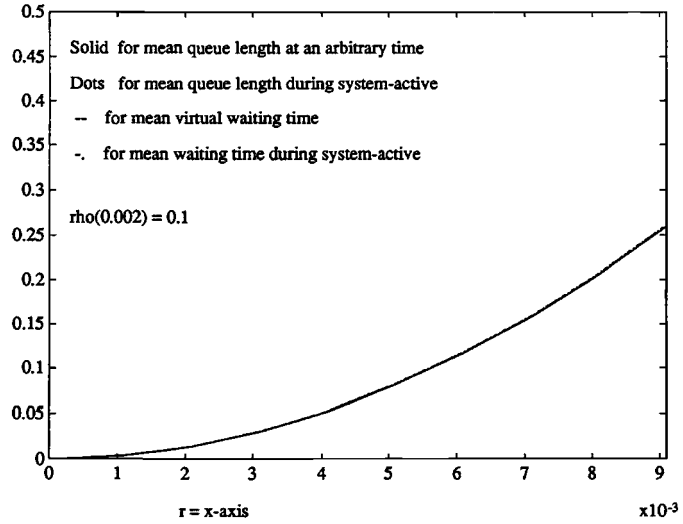


FIG. 2. *Example* 3. *Difference curves for Poisson service.*

large queue which needs to be cleared. We further noticed that the mean length of the busy period depends on the initial phase of the busy period.

Example 4. This example has a *MAP* input process with coefficient matrices

$$D_{01} = \begin{pmatrix} -2. & 0. \\ 1. & -3. \end{pmatrix}, \quad D_{02} = \begin{pmatrix} 1. & 0. \\ 0. & 1. \end{pmatrix},$$

$$D_{03} = \begin{pmatrix} 1. & 0. \\ 1. & 4. \end{pmatrix}, \quad D_{04} = \begin{pmatrix} -2. & 1. \\ 0. & -5. \end{pmatrix}, \quad D_{05} = \begin{pmatrix} 0. & 0. \\ 0. & 0. \end{pmatrix},$$

$$D_{11} = \begin{pmatrix} 0.1 & 0. \\ 0. & 0. \end{pmatrix}, \quad D_{12} = \begin{pmatrix} 0.9 & 0. \\ 0.5 & 0.5 \end{pmatrix}.$$

In contrast to Example 3, this input process is less bursty. The service processes are the same as in Example 3. See Table 2.

The difference in the magnitudes of these coefficients is striking. Only for the queue with a bursty *MAP* service process are $\mathbf{x}_{02}''\mathbf{e}$ and w_2 rather big in absolute values. The difference is caused clearly by the burstiness of the service processes. Examples 3 and 4 illustrate the influence of the burstiness of point processes on queues. In this case, Figures 3 and 4 show that the approximations are quite good for traffic intensity up to 0.15. The second-order approximation is accurate only over a slightly longer range of the traffic intensity than for the burstier input process in Example 3.

Why is this? In our experience, the accuracy of the approximation depends strongly on the Perron–Frobenius eigenvalue of the rate matrix R, which has much to do with the burstiness of the input and the service processes (see Neuts [7]).

As is to be expected, the mean number of customers served during a busy period is much smaller than for Example 3 when both cases have the same traffic intensity. However, the number of the input-active periods in a busy period grows rapidly. This

TABLE 2
Expansion coefficients of Example 4.

	Erlang	*M*	*PH*	*MMPP*	*MAP*
$\mathbf{x}'_{01}\mathbf{e}$	1.01	1.01	1.01	1.01	1.00
$\mathbf{x}''_{01}\mathbf{e}$	−1.54	−1.61	−1.61	−1.69	−5.13
$\mathbf{x}'_{02}\mathbf{e}$	−2.04	−2.04	−2.04	−2.13	−6.39
$\mathbf{x}''_{02}\mathbf{e}$	2.61	2.68	2.68	2.83	24.31
$\mathbf{q}_1\mathbf{e}$	1.06	1.06	1.06	1.15	5.56
$\mathbf{q}_2\mathbf{e}$	−0.82	−0.51	−0.49	−0.35	4.35
$\mathbf{q}_{a0}\mathbf{e}$	−0.51	0.52	0.52	0.54	0.87
$\mathbf{q}_{a1}\mathbf{e}$	0.26	0.43	0.44	0.55	3.99
w_1	0.64	1.06	1.07	1.38	30.88
w_2	−0.40	−0.51	−0.51	−0.47	−66.30
w_{a0}	0.31	0.52	0.52	0.65	4.83
w_{a1}	0.21	0.43	0.44	0.64	8.01

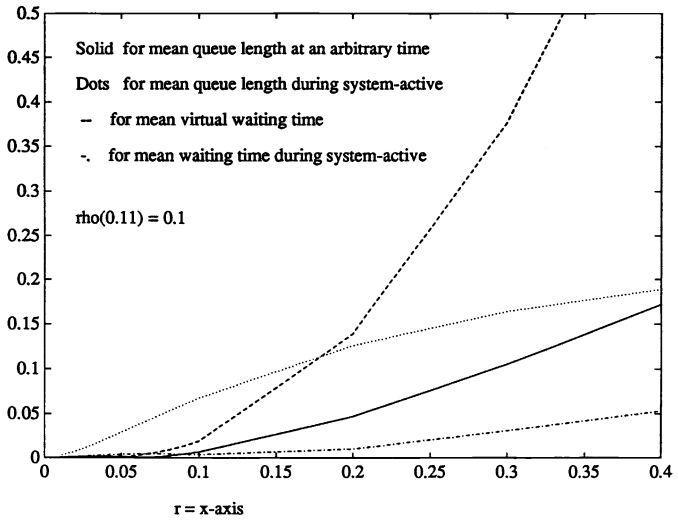


FIG. 3. *Example* 4. *Difference curves for the MAP service.*

implies that during a busy period, the input process goes to inactive states often. The number of busy periods in a system-active period is close to 2.

Appendix A. The Markovian arrival process (*MAP*). The *MAP* was first introduced in Neuts [5] as a generalization of a Poisson process. A *MAP* is defined on a finite Markov process (called the underlying Markov process) which has m states and an irreducible infinitesimal generator D. In the *MAP*, the sojourn time in state i is exponentially distributed with parameter $\lambda_i(-\geq D_{ii})$. At the end of the sojourn time in state i, a transition occurs to state j, $1 \leq j \leq m$, where the transition may or may not represent an arrival. With probability $p_{ij}(0)$, $i \neq j$, there will be a transition to state j without an arrival. With probability $p_{ij}(1)$, there will be a transition to

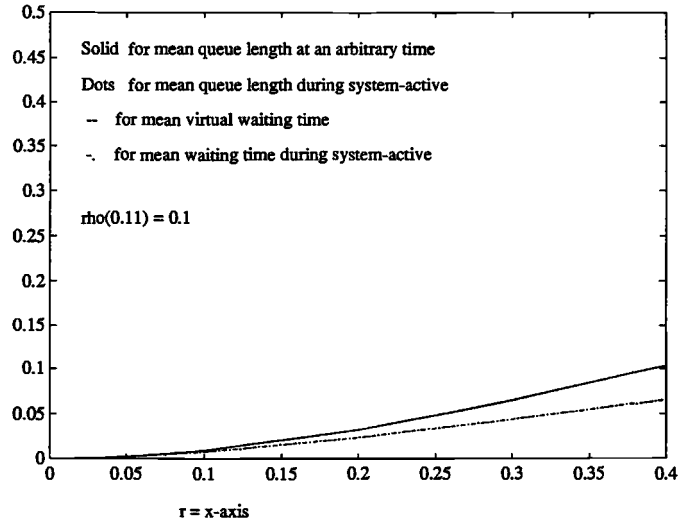


FIG. 4. *Example* 4. *Difference curves for Poisson service.*

state j with an arrival. We have

$$\sum_{j\neq i}^{m} p_{ij}(0) + \sum_{j=1}^{m} p_{ij}(1) = 1.$$

In matrix form, we denote by $(D_0)_{ii} = -\lambda_i$, $(D_0)_{ij} = \lambda_i p_{ij}(0)$, $j \neq i$ and $(D_1)_{ij} = \lambda_i p_{ij}(1)$. Then the *MAP* is represented by (D_0, D_1). D_0 is the (matrix) rate of transitions without an arrival and D_1 is the rate of transitions with an arrival. D_0 and D_1 are $m \times m$ matrices. D_0 has negative diagonal elements and nonnegative off-diagonal elements. D_1 has nonnegative elements. $D = D_0 + D_1$.

For more details about the *MAP*, see Neuts [5] and Lucantoni [4].

Appendix B. The proof of Lemma 2.1. Since $\theta_1(r)$ is a rational function of r (> 0), it is an analytic function of r. And so are $\lambda^*(r)$ and $\omega(r)$.

By (2), $\theta_{11}(r) = -r\theta_{12}(r)D_{03}(D_{01} + D_{11})^{-1}$. Then $\lim_{r\to 0} \theta_{11}(r) = 0$.

If $D_{05} = 0$, $\theta_{12}(r)Q_{22} = 0$. Then $\theta_{12}(r) = a(r)\theta_3$. By the normalization of the vector $\theta_1(r)$, we have $a(r) - a(r)r\theta_3 D_{03}(D_{01} + D_{11})^{-1}\mathbf{e} = 1$. So,

$$a(r) = \frac{1}{1 - r\theta_3 D_{03}(D_{01} + D_{11})^{-1}\mathbf{e}}.$$

So, $\lim_{r\to 0} \theta_{12}(r) = \lim_{r\to 0} a(r)\theta_3 = \theta_3$.

If $D_{05} \neq 0$, it is an irreducible infinitesimal generator. We assume that $\theta_{12}(r)$ has expansion at $r = 0$ as $\theta_{12}(r) = \theta_{12} + r\theta'_{12} + o(r)$. Substituting this into $\theta_{12}(r)(rQ_{22} + D_{05}) = 0$ and comparing the coefficients of r on both sides, we have $\theta_{12}D_{05} = 0$ and $\theta'_{12}D_{05} + \theta_{12}Q_{22} = 0$. Then $\theta_{12} = \theta_4$ and

$$\theta'_{12} = \theta_4 Q_{22}(\mathbf{e}\theta_4 - D_{05})^{-1} + (\theta'_{12}\mathbf{e})\theta_4,$$

where $\theta'_{12}\mathbf{e} = \theta_4 D_{03}(D_{01}+D_{11})^{-1}\mathbf{e}$, which is obtained by expanding the normalization equation $\theta_1(r)\mathbf{e} = 1$. Obviously, $\lim_{r\to 0} \theta_{12}(r) = \theta_4$.

The results on $\lambda^*(r)$ and $\omega(r)$ are obtained by simple calculations.

Appendix C. The proof of Theorem 3.4. Since $D(r)$ and C are irreducible and $-Z(r)\mathbf{e} < 0$, any irreducible subset of $Z(r)$ is nonconservative. Then $Z(r)$ is invertible (for $r \geq 0$). By (6), $R(r) = (D_1 \otimes I)[Z(r)]^{-1}$. By $\mathbf{X}(r)Q(r) = 0$, $\mathbf{x}_0(r)(D_1 \otimes \theta) - \mathbf{x}_1(r)Z(r) = 0$ and $\mathbf{x}_1(r) = \mathbf{x}_0(r)(I \otimes \theta)R(r)$. This gives (5). It is easy to derive the result about $\mathbf{x}_{02}$. By (5), Lemma 3.2, and Theorem 3.3, we can readily write the expansions of $\mathbf{x}_n(r)$ $(n \geq 1)$ in terms of $\mathbf{x}_0'$, $\mathbf{x}_0''$, R, R', and R''. So, we only need to expand $\mathbf{x}_0(r)$.

By Lemma 3.2, we have that $\mathbf{x}_{01}(0) = 0$. Substituting the expansion of $\mathbf{x}_0(r)$ into (5), we have

$$B1\text{: } 0 = \mathbf{x}_0'[D_0 + (I \otimes \theta)R(I \otimes (C_1\mathbf{e}))] + \mathbf{x}_0\bar{D}_0,$$
$$B2\text{: } 0 = \mathbf{x}_0''[D_0 + (I \otimes \theta)R(I \otimes C_1\mathbf{e})] + \mathbf{x}_0'[\bar{D}_0 + (I \otimes \theta)R'(I \otimes C_1\mathbf{e})].$$

More precisely, we write $B1$ and $B2$ in terms of $\mathbf{x}_{0i}$, $\mathbf{x}_{0i}'$, and $\mathbf{x}_{0i}''$, $i = 1, 2$,

$$B3\text{: } 0 = -\mathbf{x}_{01}'U_{11} + \mathbf{x}_{02}D_{03},$$
$$B4\text{: } 0 = \mathbf{x}_{01}'U_{12} + \mathbf{x}_{02}'D_{05} + \mathbf{x}_{02}D_{04},$$
$$B5\text{: } 0 = -\mathbf{x}_{01}''U_{11} + \mathbf{x}_{01}'(I \otimes \theta)R_1'(I \otimes (C_1\mathbf{e})) + \mathbf{x}_{02}'D_{03},$$
$$B6\text{: } 0 = \mathbf{x}_{01}''U_{12} + \mathbf{x}_{02}''D_{05} + \mathbf{x}_{01}'(I \otimes \theta)R_2'(I \otimes (C_1\mathbf{e})) + \mathbf{x}_{02}'D_{04},$$

where

$$U_{11} = -[D_{01} + (I \otimes \theta)R_1(I \otimes (C_1\mathbf{e}))],$$
$$U_{12} = D_{02} + (I \otimes \theta)R_2(I \otimes (C_1\mathbf{e})).$$

Let

$$U_{22} = D_{04} + D_{03}U_{11}^{-1}U_{12}.$$

U_{11} is invertible since $D(r)$ is irreducible and $-U_{11}\mathbf{e} < 0$. U_{22} is an irreducible infinitesimal generator since $U_{11}^{-1}U_{12}\mathbf{e} = \mathbf{e}$.

From $B3$, we have

$$\mathbf{x}_{01}' = \mathbf{x}_{02}D_{03}U_{11}^{-1}. \tag{52}$$

It is obvious that $\mathbf{x}_{01}' \geq 0$. By (10), $\mathbf{x}_0(r)(I \otimes \theta)(I - R(r))^{-1}\mathbf{e} = 1$. So, $\mathbf{x}_{02}(r)\mathbf{e} = 1 - (\mathbf{x}_{01}(r) \otimes \theta, 0)(I - R(r))^{-1}\mathbf{e}$. Then we have

$$\mathbf{x}_{02}'\mathbf{e} = -(\mathbf{x}_{01}' \otimes \theta, 0)(I - R)^{-1}\mathbf{e}, \tag{53}$$
$$\mathbf{x}_{02}''\mathbf{e} = -(\mathbf{x}_{01}'' \otimes \theta, 0)(I - R)^{-1}\mathbf{e} - (\mathbf{x}_{01}' \otimes \theta, 0)(I - R)^{-1}R'(I - R)^{-1}\mathbf{e}. \tag{54}$$

If $D_{05} \neq 0$, from $B4$, it is easy to obtain

$$\mathbf{x}_{02}' = \mathbf{x}_{02}U_{22}(\mathbf{e}\mathbf{x}_{02} - D_{05})^{-1} + (\mathbf{x}_{02}'\mathbf{e})\mathbf{x}_{02}. \tag{55}$$

From $B5$, we have

$$\mathbf{x}_{01}'' = [\mathbf{x}_{01}'(I \otimes \theta)R_1'(I \otimes (C_1\mathbf{e})) + \mathbf{x}_{02}'D_{03}]U_{11}^{-1}. \tag{56}$$

From $B6$, we have

$$\mathbf{x}''_{02} = [\mathbf{x}''_{01}U_{12} + \mathbf{x}'_{01}(I \otimes \theta)R'_2(I \otimes (C_1\mathbf{e})) + \mathbf{x}'_{02}D_{04}] \cdot(\mathbf{e}\mathbf{x}_{02} - D_{05})^{-1} + (\mathbf{x}''_{02}\mathbf{e})\mathbf{x}_{02}. \tag{57}$$

If $D_{05} = 0$, by $B4$, $\mathbf{x}_{02}U_{22} = 0$. So, $\mathbf{e}\mathbf{x}_{02} - U_{22}$ is invertible. By $B5$ and $B6$, we have

$$\mathbf{x}'_{02} = (\mathbf{x}'_{01} \otimes \theta)[R'_1(I \otimes (C_1\mathbf{e}))U_{11}^{-1}U_{12} + R'_2(I \otimes (C_1\mathbf{e}))] \cdot(\mathbf{e}\mathbf{x}_{02} - U_{22})^{-1} + (\mathbf{x}'_{02}\mathbf{e})\mathbf{x}_{02}. \tag{58}$$

Then $\mathbf{x}''_{01}$ is obtained from $B5$. To discover $\mathbf{x}''_{02}$, we need to expand $\mathbf{x}_0(r)$ to the third derivative and we obtain

$$\begin{aligned} B7\text{: } 0 = &-\mathbf{x}'''_{01}U_{11} + \mathbf{x}''_{01}(I \otimes \theta)R'_1(I \otimes (C_1\mathbf{e})) + \mathbf{x}''_{02}D_{03} \\ &+\mathbf{x}'_{01}(I \otimes \theta)R''_1(I \otimes (C_1\mathbf{e})), \\ B8\text{: } 0 = &-\mathbf{x}'''_{01}U_{12} + \mathbf{x}''_{01}(I \otimes \theta)R'_2(I \otimes (C_1\mathbf{e})) + \mathbf{x}''_{02}D_{04} \\ &+\mathbf{x}'_{01}(I \otimes \theta)R''_2(I \otimes (C_1\mathbf{e})). \end{aligned}$$

Similar to (55), we have

$$\mathbf{x}'''_{01} = [\mathbf{x}''_{02}D_{03} + \mathbf{x}''_{01}(I \otimes \theta)R'_1(I \otimes (C_1\mathbf{e})) + \mathbf{x}'_{01}(I \otimes \theta)R''_1(I \otimes (C_1)\mathbf{e})]U_{11}^{-1}.$$

So,

$$\begin{aligned} \mathbf{x}''_{02} = \{&\mathbf{x}''_{01}(I \otimes \theta)[R'_1(I \otimes (C_1\mathbf{e}))U_{11}^{-1}U_{12} + R'_2(I \otimes (C_1\mathbf{e}))] \\ &+\mathbf{x}'_{01}(I \otimes \theta)[R''_1(I \otimes (C_1\mathbf{e}))U_{11}^{-1}U_{12} + R''_2(I \otimes (C_1\mathbf{e}))]\} \\ &\cdot(\mathbf{e}\mathbf{x}_{02} - U_{22})^{-1} + (\mathbf{x}''_{02}\mathbf{e})\mathbf{x}_{02}. \end{aligned} \tag{59}$$

This completes the proof. Since the proof is rather complicated, the following diagram may be helpful to follow the evaluation and computation.

$$\mathbf{x}_{02}, \quad R \quad \Rightarrow \quad \mathbf{x}'_{01}.$$

For $D_{05} \neq 0$,

$$\mathbf{x}_{02}, \quad \mathbf{x}'_{01} \quad \Rightarrow \quad \mathbf{x}'_{02}\ (with\ R') \quad \Rightarrow \quad \mathbf{x}''_{01} \quad \Rightarrow \quad \mathbf{x}''_{02}.$$

For $D_{05} = 0$,

$$\mathbf{x}'_{01}, \quad R' \quad \Rightarrow \quad \mathbf{x}'_{02} \quad \Rightarrow \quad \mathbf{x}''_{01} \quad (with\ R'') \Rightarrow \quad \mathbf{x}''_{02}. \qquad \square$$

Appendix D. The proof of Lemma 4.1. If $D_{05} = 0$, the result is obvious since D_{04} is invertible. If $D_{05} \neq 0$ and is irreducible,

$$r(rD_{04} + D_{05})^{-1}D_{03} = \left(I + \frac{D_{04}^{-1}D_{05}}{r}\right)^{-1} D_{04}^{-1}D_{03}.$$

We first prove that the limit exists. Since the algebraic multiplicity of the eigenvalue 0 of D_{05} is 1, the algebraic multiplicity of the eigenvalue 0 of $D_{04}^{-1}D_{05}$ is also 1. The

geometric multiplicity of eigenvalue 0 of D_{05} is also 1 with unique positive eigenvector θ_4. Suppose that the geometric multiplicity of the eigenvalue 0 of $D_{04}^{-1}D_{05}$ is k, there is a vector $\mathbf{u} \neq 0$, $\mathbf{u}(D_{04}^{-1}D_{05})^k = 0$, and $\mathbf{u}(D_{04}^{-1}D_{05})^j \neq 0,\ 0 \leq j < k$. Then we have $\mathbf{u}(D_{04}^{-1}D_{05})^{k-1}D_{04}^{-1} = c\theta_4$, where c is a nonzero constant. If $k \geq 2$, we have $\mathbf{u}(D_{04}^{-1}D_{05})^{k-2}D_{05} = c\theta_4 D_{04}$. Postmultiplying $\mathbf{e}$ on both sides, $\theta_4 D_{04}\mathbf{e} = -\theta_4 D_{03}\mathbf{e} = 0$. This is a contradiction. Hence, $k = 1$. So, matrix $D_{04}^{-1}D_{05}$ has Jordan canonical form

$$D_{04}^{-1}D_{05} = P \begin{pmatrix} 0 & & & \\ & J_1 & & \\ & & \ddots & \\ & & & J_l \end{pmatrix} P^{-1},$$

where P is a nonsingular matrix and l is a positive integer. All the Jordan blocks $\{J_i,\ 1 \leq i \leq l\}$ have nonzero diagonal elements. Then

$$\lim_{r\to 0}\left[I + \frac{D_{04}^{-1}D_{05}}{r}\right]^{-1} = \lim_{r\to 0} P \begin{pmatrix} 1 & & & \\ & r(rI+J_1)^{-1} & & \\ & & \ddots & \\ & & & r(rI+J_l)^{-1} \end{pmatrix} P^{-1}$$

$$= P \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} P^{-1}.$$

So, the limit exists. Since the first column of the matrix P and the first row of the matrix P^{-1} are $\mathbf{e}$ and $\theta_4 D_{04}$, respectively, up to a constant scaler, we know that the limit in the above formula is $c\mathbf{e}\theta_4 D_{04}$, where c is a constant. Postmultiplying $\mathbf{e}$ on both sides of

$$\left(I + \frac{D_{04}^{-1}D_{05}}{r}\right)^{-1}\left(I + \frac{D_{04}^{-1}D_{05}}{r}\right) = I$$

and letting $r \to 0$, we obtain $c\mathbf{e}\theta_4 D_{04}\mathbf{e} = \mathbf{e}$. Hence

$$c^{-1} = \theta_4 D_{04}\mathbf{e} = -\theta_4 D_{03}\mathbf{e}$$

since $(D_{03} + D_{04})\mathbf{e} = 0$. □

REFERENCES

[1] B. Blaszczyszyn, A. Frey, and V. Schmidt, *Light-traffic approximations for Markov-modulated multi-server queues*, Stochastic Models, 11 (1995), pp. 423–445.

[2] Q.-M. He, *Differentiability of the matrices R and G in the matrix-analytic method*, Stochastic Models, 11 (1995), pp. 123–132.

[3] G.-H. Hsu and Q.-M. He, *The distribution of the first passage time of Markov processes of GI/M/1 type*, Stochastic Models, 7 (1991), pp. 397–418.

[4] D. M. Lucantoni, *New results on the single server queue with a batch Markovian arrival process*, Stochastic Models, 7 (1991), pp. 1–46.

[5] M. F. Neuts, *A versatile Markovian point process*, J. Appl. Prob., 16 (1979), pp. 764–779.

[6] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, The Johns Hopkins University Press, Baltimore, MD, 1981.
[7] M. F. Neuts, *The caudal characteristic curve of queues*, Adv. Appl. Prob., 18 (1986), pp. 221–254.
[8] M. F. Neuts, *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications*, Marcel Dekker, New York, 1989.
[9] M. F. Neuts, *The burstiness of point processes*, Stochastic Models, 9 (1993), pp. 445–466.
[10] V. Ramaswami, *The busy period of queues which have a matrix-geometric steady state probability vector*, Opsearch, 19 (1982), pp. 256–281.
[11] M. I. Reiman and B. Simon, *Light traffic limits of sojourn time distributions in Markovian queueing networks*, Stochastic Models, 4 (1988), pp. 191–234.
[12] M. I. Reiman and B. Simon, *Open queueing systems in light traffic*, Math. Oper. Res., 14 (1988), pp. 26–59.
[13] W. Van Den Hout and H. Blanc, *Development and justification of the power-series algorithm for BMAP-systems*, Stochastic Models, 11 (1995), pp. 471–496.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 249–257, January 1997

017

APPROXIMABILITY BY WEIGHTED NORMS OF THE STRUCTURED AND VOLUMETRIC SINGULAR VALUES OF A CLASS OF NONNEGATIVE MATRICES*

DANIEL HERSHKOWITZ†, WENCHAO HUANG‡, HANS SCHNEIDER‡, AND HANS WEINBERGER§

Abstract. A known result about the spectral radius of an irreducible nonnegative matrix is extended to all nonnegative matrices. By means of this result, it is shown that the structured singular value and the volumetric singular value of a class of nonnegative matrices can be approximated with arbitrary accuracy by the matrix norm induced by a weighted ℓ_2 vector norm and in the simplest case by a weighted ℓ_p vector norm for any p.

Key words. structured singular values, volumetric singular values, nonnegative matrices, weighted l_p norms

AMS subject classifications. 15A48, 15A60, 65F35

PII. S0895479895293247

1. Introduction.

We begin with a survey of some results relating the spectral radius of a matrix to the norms of matrices similar to the given matrix.

Let A be a complex $n \times n$ matrix. If $||\ ||$ is any norm on the complex n-space, the corresponding induced matrix norm $||A||^0$ is defined as the supremum of the ratio $||A\mathbf{z}||/||\mathbf{z}||$ among complex nonzero vectors $\mathbf{z}$. Because the spectral radius $\rho(A)$ is defined as the largest absolute value of the eigenvalues of A, we obtain the well-known inequality

$$\rho(A) \leq ||A||^0 \tag{1.1}$$

for any induced matrix norm by choosing $\mathbf{z}$ to be an eigenvector. It was shown by Householder ([H1, Theorem 4.4] and the remark preceding it) that $\rho(A)$ is equal to the infimum of all induced matrix norms $||A||^0$. (See also [H2, p. 46] or [HJ, Lemma 5.6.10].)

The proof consists of observing that for every $\epsilon > 0$ there is a nonsingular matrix X_ϵ such that $X_\epsilon A X_\epsilon^{-1}$ is in Jordan form with the off-diagonal elements ϵ. Householder then concludes that this implies that the matrix norm $||X_\epsilon A X_\epsilon^{-1}||_2^0$ induced by the ℓ_2 norm is arbitrarily close to $\rho(A)$, so that, in fact, $\rho(A) = \inf\|XAX^{-1}\|_2^0$.

It was shown by Bauer, Stoer, and Witzgall [BSW, Theorem 3] (see also [HJ, Theorem 5.6.37]) that the operator norm induced on a complex diagonal matrix by any absolute norm (that is, any norm whose value depends only on the absolute values of the components) is just the spectral radius of the matrix. Therefore the Householder

* Received by the editors March 29, 1995; accepted for publication (in revised form) by R. A. Horn March 19, 1996. The research of the first and third authors was supported by their joint grant 90-00434 from the United States–Israel Binational Science Foundation, Jerusalem, Israel. The research of the second and third authors was supported in part by NSF grants DMS-9123318 and DMS-9424346.

http://www.siam.org/journals/simax/18-1/29324.html

† Mathematics Department, Technion–Israel Institute of Technology, Haifa 32000, Israel (hershkowitz@tx.technion.ac.il).

‡ Mathematics Department, University of Wisconsin–Madison, Madison, WI 53706 (whuang@math.wisc.edu, hans@math.wisc.edu).

§ School of Mathematics, University of Minnesota, Vincent Hall, 206 Church St., Minneapolis, MN 55455–0487 (hfw@math.umn.edu).

argument shows that the matrix norm $||X_\epsilon A X_\epsilon^{-1}||^0$ induced by any absolute norm is arbitrarily close to $\rho(A)$. In particular, we see that for every $p \in [1, \infty]$

$$\rho(A) = \inf_{X \text{ nonsingular}} ||XAX^{-1}||_p^0, \tag{1.2}$$

where $||\ ||_p^0$ denotes the matrix norm induced by the ℓ_p norm in the complex n-space C^n. See Friedland [F] for further discussion.

For the special case of a real $n \times n$ matrix A there exists a real similarity matrix such that $X_\epsilon A X_\epsilon^{-1}$ is in a block Jordan form whose diagonal blocks are λ for real eigenvalues and $|\lambda|$ times a 2×2 orthogonal matrix for complex λ, and with the off-diagonal blocks of order ϵ. (See, e.g., [CL, Problem 40, p. 106] and the proof given in [HJ, pp. 150–153].) The above argument shows that when A is real and $p = 2$, then (1.2) still holds if the infimum is taken only over the *real* nonsingular matrices. However, an example in the appendix shows that this extension is not true, even for 2×2 matrices, when $p \neq 2$.

It was shown by Stoer and Witzgall [SW] that when A is not only real but also entrywise positive, then for any $p \in [1, \infty]$,

$$\rho(A) = \min_{X \in \mathcal{X}} ||XAX^{-1}||_p^0, \tag{1.3}$$

where

$$\mathcal{X} = \{X : X \text{ real, diagonal, and positive definite}\}. \tag{1.4}$$

This is a further improvement of (1.2) in two ways. The infimum is taken over the smaller set $\mathcal{X}$ of matrices X, and the infimum is attained. We observe that when $X \in \mathcal{X}$, the norm $||\mathbf{z}|| = ||X\mathbf{z}||_p$ which induces the matrix norm $||XAX^{-1}||_p^0$ is just a weighted ℓ_p norm.

Recently, Albrecht [A] generalized the Stoer–Witzgall result to irreducible nonnegative matrices.

We now explain the contributions of this paper. In section 2, by considering infima in place of minima, we extend the ℓ_2 case of the Albrecht–Stoer–Witzgall theorem to arbitrary nonnegative matrices. In section 3 we apply this result to produce certain classes of matrices for which the structured singular value introduced by Doyle [D] and the volumetric singular value introduced by Barmish and Polyak [BP] can be approximated with arbitrary accuracy by means of weighted ℓ_2 norms. In section 4 we show that in a special case the results in section 3 can be extended to yield sharp weighted ℓ_p bounds for the structured and the volumetric singular values.

2. Approximability of the spectral radius. If A is a square matrix with nonnegative entries, it has a nonnegative eigenvalue, the so-called Perron eigenvalue, which is equal to its spectral radius $\rho(A)$, and there exists (at least) one nonnegative left eigenvector $\mathbf{v}$ and one nonnegative right eigenvector $\mathbf{u}$ corresponding to this eigenvalue. (See, e.g., [HJ, Theorem 8.3.1].) We call such eigenvectors *Perron eigenvectors* of A. When A is also irreducible, Frobenius showed that the left and right Perron eigenvectors are unique (up to scalar multiples) and positive. (See, e.g., [HJ, Theorem 8.4.4].)

The following simple proof of a generalization which covers irreducible matrices of the special ℓ_2 case of the Stoer–Witzgall theorem is probably known, although we have been unable to find it written down.

PROPOSITION 2.1. *Let A be a matrix with nonnegative entries and suppose that there exist left and right Perron eigenvectors $\mathbf{u}$ and $\mathbf{v}$ which are positive. If*

$$X := \operatorname{diag}\{v_j^{1/2} u_j^{-1/2}\}, \tag{2.1}$$

then

$$\rho(A) = \min_{X \in \mathcal{X}} ||XAX^{-1}||_2^0 \tag{2.2}$$

so that the Stoer–Witzgall equation (1.3) *is valid when $p = 2$.*

Proof. Observe that by (2.2), $X\mathbf{u} = X^{-1}\mathbf{v}$, so that

$$(XAX^{-1})^*(XAX^{-1})X\mathbf{u} = (X^{-1}A^*X)X\rho(A)\mathbf{u} = \rho(A)^2 X^{-1}\mathbf{v} = \rho(A)^2 X\mathbf{u}.$$

Thus $\rho(A)^2$ is an eigenvalue of the nonnegative matrix $(XAX^{-1})^*(XAX^{-1})$, and the corresponding eigenvector Xu is positive. Because a positive eigenvector of a nonnegative matrix must correspond to the spectral radius (see, e.g., [HJ, Corollary 8.1.30]), the spectral radius of this matrix is $\rho(A)^2$. Thus $||XAX^{-1}||_2^0 = \rho(A)$, which is the statement of the proposition. □

We now extend the result in such a way that it also applies to all nonnegative matrices.

THEOREM 2.2. *Let A be a matrix with nonnegative entries. Then*

$$\rho(A) = \inf_{X \in \mathcal{X}} ||XAX^{-1}||_2^0, \tag{2.3}$$

where

$$\mathcal{X} = \{X : X \text{ diagonal and positive definite}\}.$$

Proof. It was shown by Frobenius that there exists a permutation matrix P such that the matrix PAP^t is block upper triangular, with the diagonal blocks irreducible. We shall assume that this permutation has been done, so that the matrix A has this form. By Proposition 2.1 there exists for each of the diagonal blocks A_{jj} a positive definite diagonal matrix X_j such that the matrix

$$D_j := X_j A_{jj} X_j^{-1}$$

has the property

$$\rho(D_j) = ||D_j||_2^0. \tag{2.4}$$

For any positive ϵ we define the block diagonal matrix

$$X_\epsilon := \operatorname{diag}\{\epsilon^{-j} X_j\}.$$

It is easily verified that

$$X_\epsilon A X_\epsilon^{-1} = D + \epsilon E, \tag{2.5}$$

where D is the block diagonal matrix with the blocks D_j, and E is a strictly upper triangular matrix whose entries are polynomials in ϵ.

Since

$$\rho(D) = \max_j \{\rho(D_j)\}$$

and

$$||D||_2^0 = \max_j \{||D_j||_2^0\},$$

the property (2.4) implies that

$$\rho(D) = ||D||_2^0. \tag{2.6}$$

We see from this, (2.5), the continuity of the norm and the spectral radius, and the definition of D that

$$\rho(A) = \rho(X_\epsilon A X_\epsilon^{-1}) = \rho(D) + o(1) = ||D||_2^0 + o(1) = ||X_\epsilon A X_\epsilon^{-1}||_2^0 + o(1).$$

In other words,

$$\lim_{\epsilon \to 0} ||X_\epsilon A X_\epsilon^{-1}||_2^0 = \rho(A).$$

Since X_ϵ is in the class $\mathcal{X}$, this implies statement (2.3) of Theorem 2.2. □

We note that a similar construction can be found in [S, p. 17] for a related problem. The example of the matrix $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ shows that the infimum in (2.3) need not be attained.

3. Structured and volumetric singular values. For any complex square matrix A the matrix norm $||A||_2^0$ induced by the ℓ_2 norm is equal to $\rho(A^*A)^{1/2}$. The latter quantity is also the largest singular value of A and is often written as $\mu(A)$.

By using the polar decomposition $A = VH$ where V is unitary and H is Hermitian positive semidefinite, one sees that $\mu(A) = \mu(H) = \rho(H)$. Since for any unitary matrix U we have $\rho(UA) \leq \mu(UA) = \mu(A)$ and since $\rho(V^*A) = \rho(H) = \mu(A)$, we see that

$$\mu(A) = \max_{U \in \mathcal{U}} \rho(UA),$$

where $\mathcal{U}$ is the group of unitary matrices.

In order to study the robustness of feedback controls, Doyle [D] introduced the concept of *structured singular value*. Let the set of n coordinate vectors of C^n be partitioned into some number ℓ of disjoint subsets. The spans of the vectors in the subsets form ℓ orthogonal subspaces which together span C^n.

For the sake of simplicity, we suppose that the indices of the vectors in any one subset are contiguous. Then an $n \times n$ matrix is naturally partitioned into a block matrix in which each block represents a transformation from one of the prescribed subspaces into the same or a different subspace.

For this reason the partition of R^n into coordinate subspaces is called a *block structure*. We shall use B to denote the block structure.

Let $\mathcal{U}_B$ be the group of unitary matrices which are block diagonal in the given block structure B. Doyle's definition of the structured singular value $\mu_s(A)$ is[1]

$$\mu_s(A) := \max_{U \in \mathcal{U}_B} \rho(UA). \tag{3.1}$$

We let $\mathcal{X}_B$ denote the set of all positive definite matrices which commute with all the matrices of $\mathcal{U}_B$. It is easily seen that

$$\begin{aligned} \mathcal{X}_B = \{X : X \quad & \text{diagonal and positive definite, and the} \\ & \text{diagonal entries of } X \text{ on each block are equal}\}. \end{aligned} \tag{3.2}$$

[1] Note that μ_s depends upon the block structure B, so that it might have been better to denote it by $\mu_B(A)$.

In connection with the same problem, Barmish and Polyak [BP, p. 8] introduced the *volumetric singular value* which may be defined as

$$\mu_v(A) := \inf_{\substack{R\in\mathcal{X}_B \\ \det(R)=1}} \mu_s(AR). \tag{3.3}$$

The infimum in (3.3) need not be attained; see [BP, p. 5].

It was observed by Doyle [D] that because the matrices of $\mathcal{X}_B$ commute with those of $\mathcal{U}_B$, because a similarity transformation leaves the spectral radius invariant, and because a unitary matrix leaves the ℓ_2 norm invariant, one finds that for any $U \in \mathcal{U}_B$ and any $X \in \mathcal{X}_B$,

$$\rho(UA) = \rho(XUAX^{-1}) = \rho(UXAX^{-1}) \le ||XAX^{-1}||_2^0. \tag{3.4}$$

By maximizing the left-hand side and minimizing the right, Doyle obtained the bound

$$\mu_s(A) \le \inf_{X\in\mathcal{X}_B} ||XAX^{-1}||_2^0. \tag{3.5}$$

By replacing A by AR, defining $Y = RX^{-1}$, and minimizing both sides of the inequality (3.4), Barmish and Polyak obtained the bound

$$\mu_v(A) \le \inf_{\substack{X,Y\in\mathcal{X}_B \\ \det(XY)=1}} ||XAY^{-1}||_2^0. \tag{3.6}$$

Because it is relatively easy to produce good algorithms to approximate the right-hand sides of (3.5) and (3.6), Doyle asked for conditions on A which ensure that equality holds in these bounds. We shall use the results of section 2 to find such conditions.

Let $\mathbf{S}_B$ denote the ℓ-dimensional subspace of R^n which consists of those vectors whose components on each of the prescribed subspaces are equal.

DEFINITION. *A matrix A is said to be* adapted *to the block structure B if $\mathbf{S}_B$ is an invariant subspace of both A and its adjoint (complex transpose) A^*.*

It is easily seen that a matrix A is adapted to the block structure B if and only if each of the ℓ^2 block matrices into which A is partitioned by the block structure has constant row sums and constant column sums.

We can now state our result on structured and volumetric singular values.

THEOREM 3.1. *Let A be nonnegative and adapted to the block structure B. Then the structured singular value of A is equal to the spectral radius of A, and equality holds in the inequalities* (3.5) *and* (3.6).

Proof. We define the nonnegative $\ell\times\ell$ matrix A_B whose ij element is the common row sum of the ij block of A. Suppose for the moment that A_B is irreducible, and let $\mathbf{u}_B$ denote its (positive) right Perron eigenvector. Define the positive n-vector $\mathbf{u} \in \mathbf{S}_B$ by requiring that its components in the jth block of B equal the jth component of $\mathbf{u}_B$. The definitions of A_B, $\mathbf{u}_B$, and $\mathbf{u}$ show that $A\mathbf{u} = \rho(A_B)\mathbf{u}$, so that $\mathbf{u}$ is a positive eigenvector of A. As in the proof of Proposition 2.1, this shows that $\rho(A) = \rho(A_B)$, so that A has the positive right Perron eigenvector $\mathbf{u}$.

Similar reasoning also shows that A also has a positive left Perron eigenvector. One must work with the matrix $\tilde{A}_B$ whose ij entry is the common column sum of the ij block, and it easy to see that $\tilde{A}_B$ is irreducible if A_B is irreducible.

We have thus shown that if A_B is irreducible, A has positive left and right Perron eigenvectors in $\mathbf{S}_B$, even though A itself may be reducible. Then Proposition 2.1 shows that if X is defined by (2.2), $\rho(A) = ||XAX^{-1}||_2^0$.

If the matrix A_B is reducible, we can put it into the Frobenius form by means of a permutation of its rows and columns. By performing the same permutations on the row and column blocks of A, we obtain a matrix which is adapted to a permutation of the partition B and whose restriction to the permuted $\mathbf{S}_B$ is in Frobenius form. We suppose this permutation to have been done beforehand, so that the matrix A_B is in the Frobenius form.

Then the matrix A is block upper triangular, with each block a union of the blocks of the structure B. The restriction of each of the diagonal blocks $A_{\alpha\alpha}$ to $\mathbf{S}_B$ is a diagonal block of the Frobenius matrix A_B and is therefore irreducible. As we showed at the beginning of this proof, this property implies that there is a positive definite diagonal matrix $X^{(\alpha)}$ which is adapted to B such that $||X^{(\alpha)}A_{\alpha\alpha}(X^{(\alpha)})^{-1}||_2^0 = \rho(A_{\alpha\alpha})$.

As in the proof of Theorem 2.2, we now define

$$X_\epsilon = \operatorname{diag}\{\epsilon^{-\alpha}X^{(\alpha)}\},$$

and show that

$$\rho(A) = \lim_{\epsilon\to 0} ||X_\epsilon A X_\epsilon^{-1}||_2^0.$$

Since $X_\epsilon \in \mathcal{X}_B$, the right-hand side is bounded below by the right-hand side of (3.5). By definition, the left-hand side is bounded above by $\mu_s(A)$ which is the left-hand side of (3.5). We conclude that $\mu_s(A) = \rho(A)$ and that the two sides of (3.5) are equal.

In order to obtain equality in (3.6), we see from the definition (3.3) that there is an $R_\epsilon \in \mathcal{X}_B$ with determinant 1 which makes $\mu_s(AR_\epsilon)$ close to $\mu_v(A)$. We use the above construction with A replaced by AR_ϵ to find an $X_\epsilon \in \mathcal{X}_B$ so that $||X_\epsilon AR_\epsilon X_\epsilon^{-1}||_2^0$ is close to $\mu_s(AR_\epsilon)$ and hence close to $\mu_v(A)$. Equality in (3.6) follows from defining $Y_\epsilon = R_\epsilon X_\epsilon^{-1}$, and the theorem is proved. □

Remark 1. The above proof shows that if, in addition to satisfying the hypotheses of Theorem 3.1, the matrix A is block irreducible in the sense that there is no nontrivial direct sum of subspaces of the partition B which is an invariant subspace of A, then there is a matrix $X \in \mathcal{X}_B$ for which $||XAX^{-1}||_2^0 = \rho(A)$.

Remark 2. Doyle [D] also considered a definition of structured singular value in which the group $\mathcal{U}_B$ in the definition (3.1) is replaced by its subgroup of real orthogonal matrices. Because Theorem 3.1 shows that $\mu_s(A) = \rho(A)$, the maximum in (3.1) is attained when $U = I$, which is real and orthogonal. Thus the statement of Theorem 3.1 is still valid for this definition.

We now observe that if A is any matrix and V and W are any matrices in $\mathcal{U}_B$, then replacing A by VAW does not change either side of (3.5) or (3.6). We can thus immediately generalize the class of matrices for which equality in (3.5) and (3.6) holds.

THEOREM 3.2. *If A is any complex matrix with the property that there exist matrices V and W in $\mathcal{U}_B$ such that VAW is nonnegative and adapted to the block structure B, then equality holds in the bounds* (3.5) *and* (3.6). □

4. Some results in ℓ_p. Additional results can be obtained for the particular block structure B_1 in which all the blocks are 1×1. In this case the set $\mathcal{U}_{B_1}$ consists of the diagonal unitary matrices, and these preserve not only the ℓ_2 norm but also all the ℓ_p norms. Therefore, we can immediately replace ℓ_2 by any ℓ_p in the bounds (3.5) and (3.6):

$$\mu_s(A) \le \inf_{X\in\mathcal{X}_{B_1}} ||XAX^{-1}||_p^0, \tag{4.1}$$

$$\mu_v(A) \leq \inf_{\substack{X,Y\in\mathcal{X}_{B_1}\\ \det(XY)=1}} ||XAY^{-1}||_p^0. \tag{4.2}$$

We note that the set $\mathcal{X}_{B_1}$ is just the set $\mathcal{X}$ of all diagonal positive definite matrices and that any matrix is adapted to the block structure B_1.

Albrecht [A] has recently obtained the following extension to nonnegative irreducible matrices of the result of Stoer and Witzgall [SW] for positive matrices. This analog of Proposition 2.1 will permit us to extend Theorem 3.1 to this case.

PROPOSITION 4.1. *Let A be an irreducible nonnegative matrix with the left and right Perron eigenvectors $\mathbf{v}$ and $\mathbf{u}$. For any $1 \leq p \leq \infty$ the matrix*

$$X = \operatorname{diag}(v_j^{1/p} u_j^{(1-p)/p}) \tag{4.3}$$

has the property that

$$\rho(A) = ||XAX^{-1}||_p^0. \qquad \square$$

By replacing Proposition 2.1 with Proposition 4.1 in the proof of Theorem 2.2, we find the following extension.

THEOREM 4.2. *If A is nonnegative and $1 \leq p \leq \infty$, then*

$$\rho(A) = \inf_{X\in\mathcal{X}} ||XAX^{-1}||_p^0. \qquad \square$$

Question: For which induced matrix norms does Theorem 4.2 hold?

By replacing Theorem 2.2 by Theorem 4.2 in the proof of Theorems 3.1 and 3.2, we immediately find the following result.

THEOREM 4.3. *Let B_1 be the block structure which consists entirely of 1×1 blocks. If the complex matrix A has the property that there are diagonal unitary matrices V and W such that the matrix VAW is nonnegative, then for any $p \in [1,\infty]$,*

$$\mu_s(A) = \inf_{X\in\mathcal{X}} ||XAX^{-1}||_p^0 = \rho(VAW)$$

and

$$\mu_v(A) = \inf_{\substack{X,Y\in\mathcal{X}\\ \det(XY)=1}} ||XAY||_p^0. \qquad \square$$

Remark 3. If C is a complex matrix, an inequality which goes back to Frobenius asserts that $\rho(C) \leq \rho(|C|)$, where $|C|$ is the matrix whose entries are the absolute values of the corresponding entries of C (see, e.g., [HJ, Theorem 8.1.18]). This inequality applied to (3.1) yields an alternative proof of the equality $\mu_s(A) = \rho(VAW)$ in Theorem 4.3.

Appendix: A counterexample. We shall show that when $p \neq 2$ and A is the 2×2 matrix $\left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$, equality does not hold in (1.2) when the infimum is taken only over real matrices X.

Any real matrix which is similar to A must have the same trace and determinant. Therefore if X is real, XAX^{-1} has the form

$$XAX^{-1} = \begin{pmatrix} 1+\alpha & -\frac{1+\alpha^2}{c} \\ c & 1-\alpha \end{pmatrix}$$

with α and c real. By inserting the two coordinate vectors into the definition of the induced matrix norm, we find that $||XAX^{-1}||_p^0$ is bounded below by the larger of the ℓ_p norms of the two columns. That is,

$$(\|XAX^{-1}\|_p^0)^p \geq \max\{|1+\alpha|^p + |c|^p,\ (1+\alpha^2)^p|c|^{-p} + |1-\alpha|^p\}.$$

We observe that for $|c| \geq |1-\alpha|$, the first expression on the right is bounded below by $|1+\alpha|^p + |1-\alpha|^p$. Since $1+\alpha^2 \geq |1+\alpha||1-\alpha|$, we see that for $|c| \leq |1-\alpha|$, the second expression on the right has the same lower bound. Thus

$$(\|XAX^{-1}\|_p^0)^p \geq |1+\alpha|^p + |1-\alpha|^p$$

for any c.

Since

$$\frac{1}{2}\begin{pmatrix}1+\alpha\\1-\alpha\end{pmatrix} + \frac{1}{2}\begin{pmatrix}1-\alpha\\1+\alpha\end{pmatrix} = \begin{pmatrix}1\\1\end{pmatrix},$$

the triangle inequality shows that $|1+\alpha|^p + |1-\alpha|^p \geq 2$. Therefore we have the lower bound

$$\|XAX^{-1}\|_p^0 \geq 2^{1/p}.$$

When $p > 2$, we use the fact that the matrix norm induced on a matrix by ℓ_p is equal to the norm induced on its conjugate transpose by the conjugate norm ℓ_q to conclude that $\inf \|XAX^{-1}\|_p^0 \geq \max\{2^{1/p}, 2^{(p-1)/p}\}$ for every p. Since $\rho(A) = 2^{1/2}$, we conclude that when $p \neq 2$, the right-hand side of (1.2) with X restricted to real matrices is strictly larger than the left-hand side.

Remark 4. By using the easily obtained equations $\|A\|_2^0 = 2^{1/2}$ and $\|A\|_1^0 = 2$ for the above matrix A and using the Riesz convexity theorem (see [R, p. 472]), one obtains the inequality $\|A\|_p^0 \leq \max\{2^{1/p}, 2^{(p-1)/p}\}$. Together with the above lower bound, this shows that for every p the minimum of $\|XAX^{-1}\|_p^0$ over real diagonal X is attained when X is the identity.

Acknowledgment. We thank Michael Neumann for comments which have helped to improve this paper.

REFERENCES

[A] J. Albrecht, *Minimal norms of non-negative, irreducible matrices*, Linear Algebra Appl., 249 (1996), pp. 255–258.

[BP] B. R. Barmish and B. T. Polyak, *The volumetric singular value and robustness of feedback control systems*, IEEE Trans. Automat. Control, to appear.

[BSW] F. L. Bauer, J. Stoer, and C. Witzgall, *Absolute and monotonic norms*, Numer. Math., 3 (1961), pp. 257–264.

[CL] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, Krieger, Melbourne, FL, 1984.

[D] J. C. Doyle, *Analysis of feedback systems with structured uncertainties*, Proc. IEEE, 129 (1982), pp. 242–250.

[F] S. Friedland, *A characterization of transform absolute norms*, Linear Algebra Appl., 28 (1979), pp. 63–68.

[H1] A. Householder, *The approximate solution of matrix problems*, J. Assoc. Comput. Mach., 5 (1958), pp. 205–243.

[H2] A. Householder, *The Theory of Matrices in Numerical Analysis*, Blaisdell, New York, 1964.

[HJ] R. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, London, 1985.
[R] M. Riesz, *Sur les maxima des forms bilinéaires sur les fonctionnelles linéaires*, Acta Math., 49 (1926), pp. 469–497.
[S] H. Schneider, *An inequality for latent roots applied to determinants with dominant principal diagonal*, J. London Math. Soc., 28 (1953), pp. 8–20.
[SW] J. Stoer and C. Witzgall, *Transformations by diagonal matrices in a normed space*, Numer. Math., 4 (1962), pp. 158–171.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, pp. 258–263, January 1997

018

SOME INEQUALITIES FOR NORMS OF COMMUTATORS*

RAJENDRA BHATIA† AND FUAD KITTANEH‡

Abstract. Let A, B be positive operators and let f be any operator monotone function. We obtain inequalities for $|||f(A)X - Xf(B)|||$ in terms of $|||f(|AX - XB|)|||$ for every unitarily invariant norm. The case $X = I$ was considered by T. Ando [*Math. Z.*, 197 (1988), pp. 403–409], and some of our results reduce to his results in this special case. Some related inequalities are obtained.

Key words. operator monotone functions, unitarily invariant norms, singular values, commutators

AMS subject classifications. 15A45, 47A30, 47A55, 47B10

PII. S0895479895293235

1. Introduction.

The aim of this paper is to present commutator versions of some perturbation inequalities proved by Ando [1] and by Jocic and Kittaneh [4]. For simplicity, we will state our results first for $n \times n$ matrices and then point out the small modifications needed to extend their validity to operators on a Hilbert space.

Let A, B be positive (semidefinite) matrices, f any nonnegative operator monotone function on $[0, \infty)$, and $|||\cdot|||$ any unitarily invariant norm. Then we have the following inequality due to Ando [1]:

$$|||f(A) - f(B)||| \leq |||f(|A - B|)|||. \tag{1}$$

Here, $|X|$ denotes $(X^*X)^{1/2}$.

Our first theorem is the following extension of this result.

THEOREM 1. *Let A, B be positive matrices. Let X be any matrix and let $s_j(X), 1 \leq j \leq n$ be the decreasingly ordered singular values of X. Then for every nonnegative operator monotone function f and for every unitarily invariant norm we have*

$$|||f(A)X - Xf(B)||| \leq \frac{1 + s_1^2(X)}{2} \left|\left|\left| f\left(\frac{2}{1 + s_n^2(X)} |AX - XB|\right)\right|\right|\right|. \tag{2}$$

After this we prove another inequality, which implies the following.

THEOREM 2. *Let A, B be positive matrices and let X be any contraction (i.e., $||X|| := s_1(X) \leq 1$). Then for every nonnegative operator monotone function f and for every unitarily invariant norm we have*

$$|||f(A)X - Xf(B)||| \leq \frac{5}{4} |||f(|AX - XB|)|||. \tag{3}$$

For the special case of the operator norm $||\cdot||$ and the power functions $f(t) = t^r,\ 0 < r \leq 1$ the inequality (3) has been proven by Pedersen [8].

Note that while the choice $X = I$ reduces the inequality (2) to (1) the same is not the case with (3). It is an interesting open question to decide whether the constant

* Received by the editors March 24, 1995; accepted for publication (in revised form) by R. A. Horn March 26, 1996.
http://www.siam.org/journals/simax/18-1/29323.html

† Indian Statistical Institute, New Delhi 110016, India (rbh@isid.ernet.in). Part of this work was done while the author was at The Fields Institute. This research was also supported by NSERC, Canada.

‡ Department of Mathematics, University of Jordan, Amman, Jordan.

5/4 occurring here could be replaced by 1. We show that for 2×2 matrices this can indeed be done.

Section 2 of this paper contains the proofs of these results, several related inequalities, and some remarks. We then obtain extensions, in the same spirit, of the following result from [4]: If A, B are Hermitian, then for every positive integer m

$$|||(A-B)^{2m+1}||| \;\le\; 2^{2m}\;|||A^{2m+1}-B^{2m+1}|||\,. \tag{4}$$

The extension we obtain is the following.

THEOREM 3. *Let A, B be Hermitian and let X be any matrix. Then for every positive integer m and for every unitarily invariant norm*

$$||| \,|AX-XB|^{2m+1}||| \;\le\; \frac{\left(1+s_1^2(X)\right)^{2m+1}}{1+s_n^2(X)}\;|||A^{2m+1}X-XB^{2m+1}|||. \tag{5}$$

If X is a contraction we have

$$||| \,|AX-XB|^{2m+1}\,||| \le 2^{2m}\left(\frac{5}{4}\right)^{2m+1}|||A^{2m+1}X-XB^{2m+1}|||. \tag{6}$$

2. Proofs and remarks. We will use standard facts about unitarily invariant norms and singular values (see, e.g., [3]) and about operator monotone functions [9]. Recall that if f is a nonnegative operator monotone function on $[0,\infty)$ then it has an integral representation

$$f(t)=\alpha+\beta t+\int_0^\infty \frac{\lambda t}{\lambda+t}d\mu(\lambda), \tag{7}$$

where $\alpha, \beta \ge 0$ and μ is a positive measure. We will repeatedly use the identity

$$f\left(UAU^*\right)=Uf(A)U^*, \tag{8}$$

valid for all unitary operators U, Hermitian operators A, and functions f whose domain contains the spectrum of A. (In the infinite-dimensional case $f(A)$ is defined via the spectral theorem for all *measurable* functions f. The representation (7) shows that operator monotone functions are infinitely differentiable.)

LEMMA 4. *For every positive A, unitary U, and nonnegative operator monotone function f on $[0,\infty)$ we have*

$$|||f(A)U-Uf(A)||| \;\le\; |||f\left(|AU-UA|\right)|||. \tag{9}$$

Proof. Using the unitary invariance of $|||\cdot|||$, the relation (8), and the inequality (1) we have

$$\begin{aligned} |||f(A)U-Uf(A)||| &= |||f(A)-Uf(A)U^*||| \\ &= |||f(A)-f\left(UAU^*\right)||| \\ &\le |||f\left(|A-UAU^*|\right)||| \\ &= |||f\left(|AU-UA|\right)|||. \end{aligned}$$

□

LEMMA 5. *Let X, Y, Z be any three matrices. Then*

$$|||f(|XYZ|)||| \;\le\; |||f\left(||X||\;||Z||\;|Y|\right)||| \tag{10}$$

for any monotone increasing function f on $[0,\infty)$.

Proof. It is an easy consequence of the min–max principle that

$$s_j(XYZ) \le ||X||\ ||Z||\ s_j(Y) \quad \text{for all } j.$$

Hence,

$$\begin{aligned} s_j\left(f(|XYZ|)\right) &= f\left(s_j(XYZ)\right) \\ &\le f\left(||X||\ ||Z||\ s_j(Y)\right) \\ &= s_j\left(f(||X||\ ||Z||\ |Y|)\right). \end{aligned}$$

This is more than adequate to ensure (10). □

The special case $A = B,\ X = X^*$. We will first prove the inequality (2) in this special case. Let

$$U = (X - i)\,(X + i)^{-1} \tag{11}$$

be the Cayley transform of X; U is unitary and its spectrum does not contain the point 1. We have

$$X = i(1+U)(1-U)^{-1} = 2i(1-U)^{-1} - i. \tag{12}$$

So, we can write

$$\begin{aligned} &|||f(A)X - Xf(A)||| \\ &\quad = \left|\left|\left|f(A)\left(2i(1-U)^{-1} - i\right) - \left(2i(1-U)^{-1} - i\right)f(A)\right|\right|\right| \\ &\quad = 2\left|\left|\left|f(A)(1-U)^{-1} - (1-U)^{-1}f(A)\right|\right|\right| \\ &\quad = 2\left|\left|\left|(1-U)^{-1}\left(f(A)U - Uf(A)\right)(1-U)^{-1}\right|\right|\right| \\ &\quad \le 2\left|\left|(1-U)^{-1}\right|\right|^2 |||f(A)U - Uf(A)||| \\ &\quad \le 2\left|\left|(1-U)^{-1}\right|\right|^2 |||f\left(|AU - UA|\right)|||, \end{aligned} \tag{13}$$

using Lemma 4. Now use (12) to obtain

$$||(1-U)^{-1}||^2 = \left|\left|\frac{X+i}{2}\right|\right|^2 = \frac{1 + s_1^2(X)}{4}. \tag{14}$$

Also note that

$$\begin{aligned} |||f\left(|AU - UA|\right)||| &= \left|\left|\left|f\left(\left|A\left(1 - 2i(X+i)^{-1}\right) - \left(1 - 2i(X+i)^{-1}\right)A\right|\right)\right|\right| \\ &= \left|\left|\left|f\left(2\left|(X+i)^{-1}A - A(X+i)^{-1}\right|\right)\right|\right| \\ &= \left|\left|\left|f\left(2\left|(X+i)^{-1}(AX - XA)(X+i)^{-1}\right|\right)\right|\right| \\ &\le \left|\left|\left|f\left(2||(X+i)^{-1}||^2\,|AX - XA|\right)\right|\right| \end{aligned} \tag{15}$$

using Lemma 5. Finally, note that

$$||(X+i)^{-1}||^2 = \frac{1}{1 + s_n^2(X)}. \tag{16}$$

The proof of (2) in the special case is completed by combining (13), (14), (15), and (16).

Proof of Theorem 1. The general case follows from the special one by a much-used trick. Let

$$C = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}, \quad Y = \begin{bmatrix} 0 & X \\ X^* & 0 \end{bmatrix}.$$

Then C is positive and Y is Hermitian. The singular values of Y are the same as those of X (but each counted twice now). The special case of the theorem applied to C in place of A and Y in place of X leads to the inequality (2). □

Proof of Theorem 2. Let t be any nonzero real number. Then the inequality (2) with tX in place of X gives

$$|||f(A)X - Xf(B)||| \leq \frac{1+t^2 s_1^2(X)}{2|t|} \left|\left|\left| f\left(\frac{2|t|}{1+t^2 s_n^2(X)} |AX - XB|\right)\right|\right|\right|. \tag{17}$$

Let $||X|| \leq 1$. Put $t = 1/2$ in (17) to get

$$|||f(A)X - Xf(B)||| \leq \frac{5}{4} \left|\left|\left| f\left(\frac{4}{4+s_n^2(X)} |AX - XB|\right)\right|\right|\right|. \tag{18}$$

Since f is operator monotone, the inequality (3) follows from (18). □

Remark 1. With slight modifications, the results above carry over to operators in an infinite-dimensional Hilbert space. We need to replace $s_1(X)$ by $||X||$ in (14) and in the subsequent discussion. In (16) we need to replace $s_n(X)$ by $\inf_{||\psi||=1} ||X\psi||$, and in the subsequent discussion we need to replace it by $\inf_{||\psi||=1} ||Y\psi||$, where

$$Y = \begin{bmatrix} X & 0 \\ 0 & X^* \end{bmatrix}.$$

Note that $\inf_{||\psi||=1} ||X\psi||$ is equal to zero if X is compact and is equal to $||X^{-1}||^{-1}$ if X is invertible.

Remark 2. In [6], Mathias showed that Ando's inequality (1) is true if f is a nonnegative *matrix monotone function of order* n on $[0, \infty)$. (This means that f is assumed to be order preserving on positive semidefinite matrices of order n only, while an *operator monotone function* is one which is matrix monotone of order n for all n.) Our proof shows that the inequalities (2) and (3) in the special case $A = B$ and $X = X^*$ are true for all functions f that are matrix monotone of order n. The proof for the general case works if f is matrix monotone of order $2n$.

Remark 3. The special case in which $f(t) = t^r,\ 0 < r \leq 1$, and the norm is the operator norm has been studied before. In [7] it was shown that for every positive A and for every X

$$||A^r X - XA^r|| \leq (1-r)^{r-1}\ ||X||^{1-r}\ ||AX - XA||^r,\ 0 < r \leq 1. \tag{19}$$

It was mentioned in that paper that Haagerup showed that the factor $(1-r)^{r-1}$ occurring in (19) could be replaced by $(\sin r\pi)/\pi r(1-r)$. This, and some extensions, were also proven in [2]. Pedersen [8], using arguments like the ones we have used, showed that the factor $(1-r)^{r-1}$ can be replaced by $5/4$. He remarks that for the special case $r = 1/2$ this can be reduced further to $2/\sqrt{\pi}$. In some special situations our inequality (2) can give better results. For example, this is so when $||X|| = 1$ and $s_n(X) > .76$.

Remark 4. For 2×2 matrices, the factor $5/4$ occurring in the inequality (3) can be replaced by 1. To see this, let

$$A = \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix}, \qquad X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}.$$

Then

$$|f(A)X - Xf(A)| = \begin{bmatrix} |(f(a_1) - f(a_2))\, x_{21}| & 0 \\ 0 & |(f(a_1) - f(a_2))\, x_{12}| \end{bmatrix}$$

and

$$f\left(|AX - XA|\right) = \begin{bmatrix} f\left(|(a_1 - a_2)\, x_{21}|\right) & 0 \\ 0 & f\left(|(a_1 - a_2)\, x_{12}|\right) \end{bmatrix}.$$

So, it is enough to show that if $|x| \le 1$, then

$$|(f(a_1) - f(a_2))\, x| \;\le\; f\left(|(a_1 - a_2)\, x|\right). \tag{20}$$

It follows from the representation (7) that $cf(t) \le f(ct)$ for $0 \le c \le 1$. So, if $x = ce^{i\theta}$, we have

$$\begin{aligned} |(f(a_1) - f(a_2))\, x| &= c\left|(f(a_1) - f(a_2))\, e^{i\theta}\right| \\ &\le cf\left(\left|(a_1 - a_2)\, e^{i\theta}\right|\right) \\ &\le f\left(|(a_1 - a_2)\, x|\right). \end{aligned}$$

Our next proposition shows that if we replace the operator norm with the Hilbert–Schmidt norm, then the first factor on the right-hand side of the inequality (19) can be replaced by 1.

PROPOSITION 6. *Let A, B be positive and let X be any matrix. Then for $0 < r < 1$,*

$$||A^r X - XB^r||_2 \;\le\; ||X||_2^{1-r}\, \|AX - XB\|_2^r. \tag{21}$$

Proof. As in Theorem 1, the general case follows from the special case $A = B$. Assume, without loss of generality, that A is diagonal with diagonal entries $\lambda_1, \ldots, \lambda_n$. Then

$$\begin{aligned} ||A^r X - XA^r||_2^2 &= \sum_{i,j} \left|\left(\lambda_i^r - \lambda_j^r\right) x_{ij}\right|^2 \\ &\le \sum_{i,j} |\lambda_i - \lambda_j|^{2r}\, |x_{ij}|^2 \\ &= \sum_{i,j} |\lambda_i - \lambda_j|^{2r}\, |x_{ij}|^{2r}\, |x_{ij}|^{2(1-r)} \\ &\le \left(\sum_{i,j} |\lambda_i - \lambda_j|^2\, |x_{ij}|^2\right)^r \left(\sum_{i,j} |x_{ij}|^2\right)^{1-r} \\ &= ||AX - XA||_2^{2r}\;\; ||X||_2^{2(1-r)}. \end{aligned}$$

We have used Hölder's inequality to arrive at our last inequality. ∎

The inequality (21) is valid for operators on Hilbert space. Let X be any Hilbert–Schmidt operator and A any positive operator. By a theorem of Weyl and von Neumann [5, p. 525] A can be expressed as a diagonal operator plus a Hilbert–Schmidt operator with arbitrarily small Hilbert–Schmidt norm. So, the same proof gives the inequality (21) in this case as well.

Following the same arguments as Ando [1] we can derive the following generalization of Theorem 2 in that paper.

THEOREM 7. *Let g be an increasing function on $[0,\infty)$ such that $g(0) = 0$, $\lim_{t\to\infty} g(t) = \infty$, and the inverse function of g is operator monotone. Then for all $A, B \geq 0$ and for all X,*

$$\frac{1+s_n^2(X)}{2}\left|\!\left|\!\left| g\left(\frac{2}{1+s_1^2(X)}\,|AX-XB|\right)\right|\!\right|\!\right| \;\leq\; |||g(A)X - Xg(B)|||. \tag{22}$$

Once again, first replacing X by tX and then making the special choice $t = 1/2$, we get from this

$$\frac{4+s_n^2(X)}{4}\left|\!\left|\!\left| g\left(\frac{4}{4+s_1^2(X)}\,|AX-XB|\right)\right|\!\right|\!\right| \leq |||g(A)X - Xg(B)|||\,. \tag{23}$$

Since g is monotonically increasing, we obtain from this the following theorem.

THEOREM 8. *Let $A, B \geq 0$ and let X be any operator with $||X|| \leq 1$. Then for every function g satisfying the conditions of Theorem 7 we have*

$$\left|\!\left|\!\left| g\left(\frac{4}{5}\,|AX-XB|\right)\right|\!\right|\!\right| \;\leq\; |||g(A)X - Xg(B)|||\,. \tag{24}$$

In particular, for every $r \geq 1$ we have

$$||| \,|AX-XB|^r||| \leq \left(\frac{5}{4}\right)^r |||A^rX - XB^r|||. \tag{25}$$

We remark that should it be possible to replace the factor $5/4$ by 1 in inequality (3), then the same could be done in (24) and (25).

The proof of Theorem 3 is analogous to that of Theorem 1. We leave the details to the reader.

Acknowledgments. We are thankful to Ken Davidson for bringing [8] to our attention.

REFERENCES

[1] T. ANDO, *Comparison of norms $|||f(A) - f(B)|||$ and $|||f(|A-B|)|||$*, Math. Z., 197 (1988), pp. 403–409.
[2] K. BOYADZHIEV, *Some inequalities for generalized commutators*, Publ. Res. Inst. Math. Sci., 26 (1990), pp. 521–527.
[3] I. C. GOHBERG AND M. G. KREIN, *Introduction to the Theory of Linear Nonselfadjoint Operators*, AMS, Providence, RI, 1969.
[4] D. JOCIC AND F. KITTANEH, *Some perturbation inequalities for self-adjoint operators*, J. Operator Theory, 31 (1994), pp. 3–10.
[5] T. KATO, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, Berlin, 1976.
[6] R. MATHIAS, *Concavity of monotone matrix functions of finite order*, Linear Multilinear Algebra, 27 (1990), pp. 129–138.
[7] C. L. OLSEN AND G. K. PEDERSEN, *Corona C^* - algebras and their applications to lifting problems*, Math. Scand., 64 (1989), pp. 63–86.
[8] G. K. PEDERSEN, *A Commutator Inequality*, preprint, Copenhagen University.
[9] M. ROSENBLUM AND J. ROVNYAK, *Hardy Classes and Operator Theory*, Oxford University Press, New York, 1985.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 1, p. 264, January 1997

019

ADDENDUM: IS THE POLAR DECOMPOSITION FINITELY COMPUTABLE?*

ALAN GEORGE† AND Kh. IKRAMOV‡

Abstract. After the paper cited in the title [George and Ikramov, *SIAM J. Matrix Anal. Appl.*, 17 (1996), pp. 348–354] was sent to the printer, the authors succeeded in finding a way to show that the answer to the title question is no. This brief note contains a proof of the result.

Key words. polar decomposition, finite computation

AMS subject classification. 65F10

PII. S0895479896303260

Assume that, for any A, the polar decomposition of A can be computed finitely. In particular, for a Hermitian matrix A, it is possible to compute finitely the positive semidefinite matrix $H = (A^2)^{1/2}$. One can assume, without loss of generality, that A has a simple spectrum. Otherwise, we could first transform A to tridiagonal form and deal afterwards with the irreducible diagonal blocks in the latter matrix. In what follows, it is important to note that since the spectrum is simple, it is possible, using a finite amount of arithmetic, to find a shift γ such that $A - \gamma I$ has a single positive eigenvalue.

Let

$$C = (A + H)/2.$$

Then the column space and the null space of C coincide with the invariant subspaces of A corresponding to its positive and nonpositive eigenvalues, respectively. It is obvious that a basis of the column space and that of the null space of any matrix can be computed finitely (and even rationally). (We mention that, for our problem, orthonormal bases of both subspaces are given by columns of the unitary factor in the polar decomposition of A.) It follows that, for a Hermitian A with a single positive eigenvalue λ, the eigenspace L associated with λ can be found finitely. Then any nonzero vector $x \in L$ can be used to find λ via the relation $Ax = \lambda x$. Since this argument is applicable to any shifted matrix $A - \alpha I$, we conclude that the spectrum of any Hermitian matrix can be computed finitely. It is widely known that this is not true, in general, for $n \geq 5$.

REFERENCE

[1] A. GEORGE AND Kh. IKRAMOV, *Is the polar decomposition finitely computable?*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 348–354.

* Received by the editors April 15, 1996; accepted for publication by C. Van Loan May 9, 1996. This work was supported by Natural Sciences and Engineering Research Council of Canada grant OGP 000811.

http://www.siam.org/journals/simax/18-1/30326.html

† Department of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada (jageorge@sparse1.uwaterloo.ca).

‡ Faculty of Numerical Mathematics and Cybernetics, Moscow State University, Moscow 119899, Russia (ikramov@cmc.msk.su).

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, pp. 265–278, April 1997

001

ON PARABOLIC AND ELLIPTIC SPECTRAL DICHOTOMY*

A. N. MALYSHEV† AND M. SADKANE‡

Abstract. We discuss two spectral dichotomy techniques: one for computing an invariant subspace of a nonsymmetric matrix associated with the eigenvalues inside and outside a given parabola; another one for computing a right deflating subspace of a regular matrix pencil associated with the eigenvalues inside and outside a given ellipse. The techniques use matrices of twice the order of the original matrices on which the spectral dichotomy by the unit circle and by the imaginary axis apply efficiently. We prove the equivalence between the condition number of the original problems and that of the transformed ones.

Key words. eigenvalue, regular matrix pencil, spectral dichotomy, spectral transformation, spectral condition number

AMS subject classifications. 65F15, 30C20

PII. S0895479894277958

1. Introduction.

In this note we are concerned with some spectral transformations for computing right eigenspaces corresponding to the eigenvalues of a matrix (matrix pencil) in a given domain $\mathcal{D}$ of the complex plane.

We first consider the case where $\mathcal{D}$ is the interior (exterior) of a given parabola γ and where we seek a right eigenspace of a matrix A having no eigenvalues on γ. By using a special matrix $\mathcal{A}$, we show that we reduce the problem to that of the computation of a right eigenspace of $\mathcal{A}$ corresponding to the left (right) half plane of the complex plane. Therefore, the results concerning the spectral dichotomy by the imaginary axis [11, 19, 20, 21] or matrix sign function [18, 7, 17] can be applied efficiently.

In the second part we consider the case where $\mathcal{D}$ is the interior (exterior) of an ellipse Γ and where we seek a right eigenspace of a regular matrix pencil $\lambda B - A$ having no eigenvalues on Γ. By using a special regular matrix pencil $\lambda\mathcal{B} - \mathcal{A}$, we show that this problem is reduced to that of the computation of a right eigenspace of the pencil $\lambda\mathcal{B} - \mathcal{A}$ corresponding to the interior (exterior) of the unit circle. Therefore, the results concerning the circular spectral dichotomy [5, 19, 20, 21] can be applied efficiently. This latter problem has recently been studied in [12]. We propose here a simple solution based on standard linear algebra without making use of Green matrices.

We study the condition numbers of these two problems and prove the equivalence between them and those of the transformed ones.

We mention that other possible approaches for solving these types of problems include the inverse free spectral divide-and-conquer methods [2, 3] and the Schur-based methods [8, 9, 14, 15].

Throughout this note, we use some standard conformal mappings that may be found, for example, in [1]. The symbol $\| x \|$ denotes the Euclidean norm of the vector x. $\| X \|$ and X^* denote, respectively, the spectral norm and the conjugate transpose of the matrix X. We denote by I_n the identity matrix of order n.

* Received by the editors November 30, 1994; accepted for publication (in revised form) by B. Kågström March 6, 1996.
http://www.siam.org/journals/simax/18-2/27795.html
† Institute of Mathematics, Novosibirsk, 630090, Russia (malyshev@math.nsk.su). This work was carried out while the author was visiting IRISA. The work was supported by DRET and partially by the Russian Fund of Fundamental Researches (93-011-1515).
‡ IRISA-INRIA, Campus de Beaulieu, 35042 Rennes Cedex, France (sadkane@irisa.fr).

2. Parabolic dichotomy. Let us consider a parabola γ in the (x, y) plane satisfying the equation

$$y^2 = 2p(p/2 - x), \tag{1}$$

where $p > 0$ is some real parameter. This parabola has its center at the origin and its branches go to the left half plane symmetrically with respect to the real axis.

Consider the mapping $\varphi : \lambda \in \mathbb{C} \mapsto z \in \mathbb{C}$ defined by the formula $z = (\lambda + \sqrt{p/2})^2$. If we write $z = x + iy$ then we obtain

$$x = \left(\Re\lambda + \sqrt{p/2}\right)^2 - (\Im\lambda)^2 \quad \text{and} \quad y = 2\left(\Re\lambda + \sqrt{p/2}\right)\Im\lambda. \tag{2}$$

It follows that $y^2 = 4(\Re\lambda + \sqrt{p/2})^2[(\Re\lambda + \sqrt{p/2})^2 - x]$. Thus φ bijectively maps the straight line $\Re\lambda = c \in \mathbf{R}$ onto the parabola $y^2 = 2\bar{p}(\bar{p}/2 - x)$ with $\bar{p} = 2(c + \sqrt{p/2})^2$. This family of parabolae depending on the parameter $\bar{p}$ has the same center at the origin. In particular, φ bijectively maps the imaginary axis onto the parabola γ. It is easy to see that φ maps the region

$$\{\lambda \in \mathbb{C} |\ \Re\lambda < -2\sqrt{p/2}\ \} \cup \{\lambda \in \mathbb{C} |\ \Re\lambda > 0\ \}$$

onto the exterior of γ and maps the strip

$$\{\lambda \in \mathbb{C} |\ -2\sqrt{p/2}\ < \Re\lambda < 0\ \}$$

onto the interior of γ. The restriction of the mapping φ on the set

$$\Omega = \{\lambda \in \mathbb{C} |\ \Re\lambda > -\sqrt{p/2}\ \}$$

is one-to-one and $\varphi(\Omega) = \{\lambda \in \mathbb{C} |\ (\Re\lambda > 0)\, or\, (\Im\lambda \neq 0)\ \}$; that is, $\varphi(\Omega)$ is the whole (x, y) plane with the real negative axis deleted. Finally, it is easy to see that the right half plane, i.e., the set of λ with $\Re\lambda > 0$, is mapped conformally onto the exterior of the parabola $y^2 = 2p(p/2 - x)$.

After these preliminaries we now consider a matrix A of order n having no eigenvalues on the parabola γ of equation $y^2 = 2p(p/2 - x)$. Let $\mathcal{A}$ be the matrix of order $2n$ defined by

$$\mathcal{A} = \begin{bmatrix} -\sqrt{\frac{p}{2}} I_n & A \\ I_n & -\sqrt{\frac{p}{2}} I_n \end{bmatrix}. \tag{3}$$

It is easy to see that the eigenvalues λ of $\mathcal{A}$ and the eigenvalues z of A satisfy

$$z = \left(\lambda + \sqrt{p/2}\right)^2 = \varphi(\lambda). \tag{4}$$

From the above discussion about the properties of the mapping φ and the assumption about the eigenvalues of A, we see that the matrix $\mathcal{A}$ has no eigenvalues on the imaginary axis. Thus the problem of verification of absence of the eigenvalues of the matrix A on the parabola γ is transformed into the spectral dichotomy problem with respect to the imaginary axis. This latter problem has been deeply studied from both theoretical and practical aspects [11, 20, 21].

We assume from now on that $||A|| = 1$. This assumption is not restrictive since otherwise we can take $A_1 = \frac{1}{||A||} A$ and $p_1 = \frac{p}{||A||}$.

In the spectrum dichotomy problem for the matrix $\mathcal{A}$ with respect to the imaginary axis, the quality of the dichotomy is characterized by the numerical parameter [20, 2]

$$\tilde{\alpha} = \sup_{\Re\lambda=0} \|(\lambda I_{2n} - \mathcal{A})^{-1}\|. \tag{5}$$

This parameter has already been used in the context of matrix stability [25, 6]. Similarly, the quality of the dichotomy for the matrix A with respect to the parabola γ is characterized by the parameter

$$\alpha = \sup_{z\in\gamma} \|(zI_n - A)^{-1}\|. \tag{6}$$

The natural question that one may ask is how to relate the two quantities α and $\tilde{\alpha}$. The answer is given in the following proposition.

PROPOSITION 2.1. *Let α and $\tilde{\alpha}$ be the two parameters defined in* (5) *and* (6). *Then*

$$\alpha \le \tilde{\alpha} \le \alpha + \sqrt{\alpha}\sqrt{1+\alpha}. \tag{7}$$

Proof. We have

$$(\lambda I_{2n} - \mathcal{A})^{-1} = \begin{bmatrix} \left(\lambda + \sqrt{\frac{p}{2}}\right) I_n & A \\ I_n & \left(\lambda + \sqrt{\frac{p}{2}}\right) I_n \end{bmatrix} \times \begin{bmatrix} \left(\lambda + \sqrt{\frac{p}{2}}\right)^2 I_n - A & 0 \\ 0 & \left(\lambda + \sqrt{\frac{p}{2}}\right)^2 I_n - A \end{bmatrix}^{-1}.$$

By considering the $n \times n$ block in position $(2,1)$ of the resulting matrix, we obtain

$$\tilde{\alpha} = \sup_{\Re\lambda=0} \|(\lambda I_{2n} - \mathcal{A})^{-1}\| \ge \sup_{z\in\gamma} \|(zI_n - A)^{-1}\| = \alpha.$$

The use of (4) gives

$$\|(\lambda I_{2n} - \mathcal{A})^{-1}\| \le \left\| \begin{pmatrix} \sqrt{z} I_n & A \\ I_n & \sqrt{z} I_n \end{pmatrix} \right\| \|(zI_n - A)^{-1}\|,$$

and since $||A|| = 1$ we have

$$\|(\lambda I_{2n} - \mathcal{A})^{-1}\| \le \|(zI_n - A)^{-1}\|(1 + \sqrt{|z|}).$$

We deduce that if $|z| \le \frac{1+\alpha}{\alpha}$, then $||(\lambda I_{2n} - \mathcal{A})^{-1}|| \le \alpha(\sqrt{\frac{1+\alpha}{\alpha}} + 1) = \alpha + \sqrt{\alpha}\sqrt{1+\alpha}$. Now if $|z| > \frac{1+\alpha}{\alpha}$, the formula $(zI_n - A)^{-1} = \frac{1}{z} I_n + \frac{1}{z} A(zI_n - A)^{-1}$ yields

$$\begin{aligned} ||(\lambda I_{2n} - \mathcal{A})^{-1}|| &\le \left(||(zI_n - A)^{-1}|| + 1\right) \frac{\sqrt{|z|} + 1}{|z|} \\ &\le (\alpha + 1)\left(\frac{1}{|z|} + \frac{1}{\sqrt{|z|}}\right) < \alpha + \sqrt{\alpha}\sqrt{1+\alpha}. \end{aligned}$$

∎

Thus, provided that α is not small, the quality of the spectral dichotomy of the matrix A with respect to the parabola γ and that of the matrix $\mathcal{A}$ with respect to the imaginary axis is equivalent.

We now describe how to obtain a solution to the spectral dichotomy problem for A with respect to the parabola γ, having calculated a solution to the spectral dichotomy problem for $\mathcal{A}$ with respect to the imaginary axis. Let $P \in \mathbf{C^{n\times n}}$ be the projection matrix onto the right eigenspace of A associated with the eigenvalues outside the parabola γ and $\mathcal{P} \in \mathbf{C^{2n\times 2n}}$ the projection matrix onto the right eigenspace of $\mathcal{A}$ associated with the eigenvalues on the right half plane of the complex plane. The following proposition characterizes the relation between P and $\mathcal{P}$.

PROPOSITION 2.2. *Let us partition $\mathcal{P}$ in the following form:*

$$\mathcal{P} = \begin{pmatrix} \mathcal{P}_1 & \mathcal{P}_2 \\ \mathcal{P}_3 & \mathcal{P}_4 \end{pmatrix} \quad \textit{with} \quad \mathcal{P}_i \in \mathbf{C^{n\times n}}, \quad i = 1, 4. \tag{8}$$

Then

$$P = 2\mathcal{P}_1 = 2\mathcal{P}_4 = 4\mathcal{P}_2\mathcal{P}_3. \tag{9}$$

Moreover,

$$\mathcal{P}_2 = \frac{1}{2}(PA)^{\frac{1}{2}}. \tag{10}$$

Proof. Suppose that the matrix A is nonsingular (below this restriction will be removed). Let us define a solution X to the matrix equation

$$\left(X + \sqrt{\frac{p}{2}} I_n\right)^2 = A, \tag{11}$$

whose eigenvalues are assumed to be in the domain

$$\mathcal{D} = \{\lambda \in \mathbb{C} |\ \Re\lambda > -\sqrt{p/2}\ \} \cup \{\lambda \in \mathbb{C} |\ \Re\lambda = -\sqrt{p/2} \text{ and } \Im\lambda > 0\ \}.$$

It is clear that such a matrix X exists and is uniquely defined.[1] Notice that the matrix $X_1 = -X - 2\sqrt{p/2}\, I_n$ also satisfies the equation $(X_1 + \sqrt{p/2}\, I_n)^2 = A$.

The matrix $\mathcal{A}$ can thus be decomposed in the following form:

$$\begin{aligned} \mathcal{A} &= \begin{pmatrix} X + \sqrt{\frac{p}{2}}\, I_n & -X - \sqrt{\frac{p}{2}}\, I_n \\ I_n & I_n \end{pmatrix} \begin{pmatrix} X & 0 \\ 0 & -X - 2\sqrt{\frac{p}{2}}\, I_n \end{pmatrix} \\ &\quad \times \begin{pmatrix} X + \sqrt{\frac{p}{2}}\, I_n & -X - \sqrt{\frac{p}{2}}\, I_n \\ I_n & I_n \end{pmatrix}^{-1} \end{aligned} \tag{12}$$

$$\begin{aligned} &= \begin{pmatrix} X + \sqrt{\frac{p}{2}}\, I_n & -X - \sqrt{\frac{p}{2}}\, I_n \\ I_n & I_n \end{pmatrix} \begin{pmatrix} X & 0 \\ 0 & -X - 2\sqrt{\frac{p}{2}}\, I_n \end{pmatrix} \\ &\quad \times \frac{1}{2} \begin{pmatrix} \left(X + \sqrt{\frac{p}{2}}\, I_n\right)^{-1} & I_n \\ -\left(X + \sqrt{\frac{p}{2}}\, I_n\right)^{-1} & I_n \end{pmatrix}. \end{aligned} \tag{13}$$

Consider the Jordan canonical form of the matrix X:

$$X = Q \begin{pmatrix} J_+ & 0 \\ 0 & J_- \end{pmatrix} Q^{-1}, \tag{14}$$

[1] The properties of the mapping φ ensure the uniqueness of X in $\mathcal{D}$.

where J_+ (J_-) corresponds to the Jordan block associated with the eigenvalues of X in the right (left) half plane. Using the expression (14) of X, the decomposition (12) of $\mathcal{A}$ reduces to

$$\mathcal{A} = \mathcal{Q}\mathcal{J}\mathcal{Q}^{-1}, \tag{15}$$

where

$$\mathcal{Q} = \begin{pmatrix} Q & 0 \\ 0 & Q \end{pmatrix} \begin{bmatrix} \begin{pmatrix} J_+ & 0 \\ 0 & J_- \end{pmatrix} + \sqrt{\tfrac{p}{2}} I_n & -\begin{pmatrix} J_+ & 0 \\ 0 & J_- \end{pmatrix} - \sqrt{\tfrac{p}{2}} I_n \\ I_n & I_n \end{bmatrix}$$

and

$$\mathcal{J} = \begin{bmatrix} \begin{pmatrix} J_+ & 0 \\ 0 & J_- \end{pmatrix} & 0 \\ 0 & -\begin{pmatrix} J_+ & 0 \\ 0 & J_- \end{pmatrix} - 2\sqrt{\tfrac{p}{2}} I_n \end{bmatrix}.$$

Notice that the expressions of $\mathcal{P}$ and P are simply

$$\mathcal{P} = \mathcal{Q} \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} \mathcal{Q}^{-1} \quad \text{and} \quad P = Q \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} Q^{-1}, \tag{16}$$

where k is the order of the matrix J_+. The expression of $\mathcal{P}$ in (16) can easily be written in form (8) with

$$\mathcal{P}_1 = \mathcal{P}_4 = Q \begin{pmatrix} \frac{1}{2} I_k & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} \equiv \frac{1}{2} P, \tag{17}$$

$$\mathcal{P}_2 = Q \begin{pmatrix} \frac{1}{2}(J_+ + \sqrt{\frac{p}{2}} I_k) & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} \equiv \frac{1}{2} (PA)^{\frac{1}{2}}, \tag{18}$$

and

$$\mathcal{P}_3 = Q \begin{pmatrix} \frac{1}{2}(J_+ + \sqrt{\frac{p}{2}} I_k)^{-1} I_n & 0 \\ 0 & 0 \end{pmatrix} Q^{-1}. \tag{19}$$

It remains only to observe that by the continuity arguments, that is, by considering, for example, the matrix $A_\epsilon = A + \epsilon I_n$, one can remove the nonsingularity assumption on A. □

3. Elliptic dichotomy. Let $\lambda B - A$ be a regular matrix pencil of order n having no eigenvalues on the ellipse Γ of equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1. \tag{20}$$

We assume throughout this section that $a \geq b > 0$. Consider the mapping $\psi : \xi \in \mathbb{C} \mapsto \lambda \in \mathbb{C}$ defined by the formula

$$\lambda = \psi(\xi) = \frac{(a+b)\xi^2 + (a-b)}{2\xi}. \tag{21}$$

It is easy to see that ψ maps the region

$$R_1 = \left\{ \xi \in \mathbb{C} |\ |\xi| < \frac{a-b}{a+b} \right\} \cup \{ \xi \in \mathbb{C} |\ |\xi| > 1 \}$$

onto the exterior of Γ and maps the annulus

$$R_2 = \left\{ \xi \in \mathbb{C} |\ \frac{a-b}{a+b} < |\xi| < 1 \right\}$$

onto the interior of Γ. In particular, ψ conformally transforms the exterior of the unit circle onto the exterior of the ellipse Γ. The mapping ψ is a bijection when $r > \sqrt{\frac{a-b}{a+b}}$.

Now consider the quadratic matrix pencil

$$\frac{a+b}{2} B\xi^2 - A\xi + \frac{a-b}{2} B. \tag{22}$$

Then

$$\frac{a+b}{2} B\xi^2 - A\xi + \frac{a-b}{2} B = \xi(\lambda B - A), \tag{23}$$

where $\lambda = \psi(\xi)$ is defined in (21). This shows the relationship between the eigenstructure of the quadratic matrix pencil (22) and that of the original pencil $\lambda B - A$.

From (23) and the assumption on the eigenvalues of $\lambda B - A$, we see that the quadratic matrix pencil (22) has no eigenvalues on the unit circle and that

$$\sup_{|\xi|=1} \left\| \left[\frac{a+b}{2} B\xi^2 - A\xi + \frac{a-b}{2} B \right]^{-1} \right\| = \sup_{\lambda \in \Gamma} \|(\lambda B - A)^{-1}\|. \tag{24}$$

A classical way of dealing with the quadratic matrix pencil (22) is to consider a matrix pencil [10] of the form $\mu\mathcal{B} - \mathcal{A}$, where

$$\mathcal{B} = \begin{pmatrix} \frac{a+b}{2} B & -A \\ 0 & \frac{a+b}{2} B \end{pmatrix} \quad \text{and} \quad \mathcal{A} = \begin{pmatrix} -\frac{a-b}{2} B & 0 \\ A & -\frac{a-b}{2} B \end{pmatrix}.$$

It is easy to show that if $\left(\frac{a+b}{2} B\xi^2 - A\xi + \frac{a-b}{2} B\right) x = 0$, then $\left(\xi^2 \mathcal{B} - \mathcal{A}\right) \left(\begin{smallmatrix} \xi x \\ x \end{smallmatrix}\right) = 0$. From the assumption on the eigenvalues of $\lambda B - A$, we see that the eigenvalues of the matrix pencil $\mu\mathcal{B} - \mathcal{A}$ cannot be on the unit circle. Thus the problem of verification of absence of the eigenvalues of the matrix pencil $\lambda B - A$ on the ellipse Γ is reduced to a spectral dichotomy problem with respect to the unit circle; this latter problem is now well understood [5, 19, 20, 21] and can be applied efficiently.

In the spectrum dichotomy problem for the matrix pencil $\lambda B - A$, the quality of the dichotomy is characterized by the numerical parameter [20, 21, 2]

$$\beta = \sup_{\lambda \in \Gamma} \|(\lambda B - A)^{-1}\|. \tag{25}$$

Similarly, the quality of the dichotomy for the matrix pencil $\mu\mathcal{B} - \mathcal{A}$ with respect to the unit circle is characterized by the numerical parameter

$$\tilde{\beta} = \sup_{|\mu|=1} \|(\mu\mathcal{B} - \mathcal{A})^{-1}\|. \tag{26}$$

As in the previous section, we begin by comparing the two quantities β and $\tilde{\beta}$.

PROPOSITION 3.1. *The quantities β and $\tilde{\beta}$ defined in* (25) *and* (26) *are equal:*

$$\beta = \tilde{\beta}. \tag{27}$$

Proof. We have

$$\begin{aligned}
(\mu^2\mathcal{B}-\mathcal{A})^{-1} &= \begin{pmatrix} \lambda\mu B & -A\mu^2 \\ -A & \lambda\mu B \end{pmatrix}^{-1} \quad \text{with} \quad \lambda=\psi(\mu) \\
&= \begin{pmatrix} I_n & 0 \\ 0 & \mu^{-1}I_n \end{pmatrix} \begin{pmatrix} \lambda B & -A \\ -A & \lambda B \end{pmatrix}^{-1} \begin{pmatrix} \mu^{-1}I_n & 0 \\ 0 & I_n \end{pmatrix} \\
&= \begin{pmatrix} I_n & 0 \\ 0 & \mu^{-1}I_n \end{pmatrix} \\
&\times \frac{1}{2} \begin{bmatrix} (\lambda B-A)^{-1}+(\lambda B+A)^{-1} & (\lambda B-A)^{-1}-(\lambda B+A)^{-1} \\ (\lambda B-A)^{-1}-(\lambda B+A)^{-1} & (\lambda B-A)^{-1}+(\lambda B+A)^{-1} \end{bmatrix} \\
&\times \begin{pmatrix} \mu^{-1}I_n & 0 \\ 0 & I_n \end{pmatrix} \\
&= \begin{pmatrix} I_n & 0 \\ 0 & \mu^{-1}I_n \end{pmatrix} \begin{pmatrix} \frac{I_n}{\sqrt{2}} & -\frac{I_n}{\sqrt{2}} \\ \frac{I_n}{\sqrt{2}} & \frac{I_n}{\sqrt{2}} \end{pmatrix} \begin{bmatrix} (\lambda B-A)^{-1} & 0 \\ 0 & (\lambda B+A)^{-1} \end{bmatrix} \\
&\times \begin{pmatrix} \frac{I_n}{\sqrt{2}} & \frac{I_n}{\sqrt{2}} \\ -\frac{I_n}{\sqrt{2}} & \frac{I_n}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \mu^{-1}I_n & 0 \\ 0 & I_n \end{pmatrix}. \qquad \square
\end{aligned}$$

We now describe how to obtain a solution to the spectral dichotomy problem for the pencil $\lambda B - A$ with respect to the ellipse Γ, having calculated a solution to the spectral dichotomy problem for the pencil $\mu\mathcal{B}-\mathcal{A}$ with respect to the unit circle. Let $P \in \mathbf{C}^{\mathbf{n}\times\mathbf{n}}$ be the projection matrix onto the right eigenspace of the pencil $\lambda B - A$ associated with the eigenvalues outside the ellipse Γ and $\mathcal{P} \in \mathbf{C}^{\mathbf{2n}\times\mathbf{2n}}$ the projection matrix onto the right eigenspace of $\mu\mathcal{B}-\mathcal{A}$ associated with the eigenvalues outside the unit circle. The following proposition characterizes the relation between P and $\mathcal{P}$.

PROPOSITION 3.2. *Let us partition $\mathcal{P}$ in the following form:*

$$\mathcal{P} = \begin{pmatrix} \mathcal{P}_1 & \mathcal{P}_2 \\ \mathcal{P}_3 & \mathcal{P}_4 \end{pmatrix} \quad \textit{with} \quad \mathcal{P}_i \in \mathbf{C}^{\mathbf{n}\times\mathbf{n}}, \quad i=1,4. \tag{28}$$

Then

$$P = \mathcal{P}_1 + \mathcal{P}_4. \tag{29}$$

Proof. Assume that the pencil $\lambda B - A$ has no infinite eigenvalues (this restriction is removed by the continuity arguments afterwards). Let us define a solution X to the matrix equation

$$\frac{a+b}{2}BX^2 - AX + \frac{a-b}{2}B = 0, \tag{30}$$

whose eigenvalues are assumed to be outside the circle of radius $\frac{a-b}{a+b}$. Since the eigenvalues ξ of X are chosen by formula (21), that is, $\xi = \frac{\lambda+\sqrt{\lambda^2-a^2+b^2}}{a+b}$, where

$\lambda = \psi(\xi)$ is an eigenvalue of the pencil $\lambda B - A$, the eigenvalues of X corresponding to λ outside the ellipse Γ are uniquely defined.[2]

It is easy to see that the matrix $X_1 = \frac{a-b}{a+b}X^{-1}$ also satisfies $\frac{a+b}{2}BX_1^2 - AX_1 + \frac{a-b}{2}B = 0$.

Consider the Jordan canonical form of the matrix X:

$$X = Q \begin{pmatrix} J_\infty & 0 \\ 0 & J_0 \end{pmatrix} Q^{-1}, \tag{31}$$

where J_∞ (J_0) corresponds to the Jordan block associated with the eigenvalues of X outside (inside) the unit circle. From (31) and the identity

$$\begin{bmatrix} \frac{a-b}{2}B & 0 \\ -A & \frac{a-b}{2}B \end{bmatrix} \begin{bmatrix} X & X_1 \\ I_n & I_n \end{bmatrix} + \begin{bmatrix} \frac{a+b}{2}B & -A \\ 0 & \frac{a+b}{2}B \end{bmatrix} \begin{bmatrix} X & X_1 \\ I_n & I_n \end{bmatrix} \begin{bmatrix} X^2 & 0 \\ 0 & X_1^2 \end{bmatrix} = 0$$

we see that

$$\mathcal{P} = \begin{pmatrix} X & X_1 \\ I_n & I_n \end{pmatrix} \begin{bmatrix} Q \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} X & X_1 \\ I_n & I_n \end{pmatrix}^{-1},$$

where k is the order of the matrix J_∞. Note that the matrix $X - X_1$ is nonsingular if and only if the matrix pencil $\lambda B - A$ does not have $\pm\sqrt{a^2 - b^2}$ as eigenvalues. This can be assumed without loss of generality (continuity arguments). Thus

$$\begin{aligned}
\mathcal{P} &= \begin{pmatrix} X & X_1 \\ I_n & I_n \end{pmatrix} \begin{bmatrix} Q \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} I_n & -X_1 \\ -I_n & X \end{pmatrix} \begin{pmatrix} X - X_1 & 0 \\ 0 & X - X_1 \end{pmatrix}^{-1} \\
&= \begin{pmatrix} Q & 0 \\ 0 & Q \end{pmatrix} \begin{bmatrix} J_\infty & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ I_k & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} I_k & 0 & -\frac{a-b}{a+b}J_\infty^{-1} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} Q & 0 \\ 0 & Q \end{pmatrix}^{-1} \\
&\quad \times \begin{pmatrix} X - X_1 & 0 \\ 0 & X - X_1 \end{pmatrix}^{-1} \\
&= \begin{bmatrix} Q \begin{pmatrix} J_\infty & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} & Q \begin{pmatrix} -\frac{a-b}{a+b}I_k & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} \\ Q \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} & Q \begin{pmatrix} -\frac{a-b}{a+b}J_\infty^{-1} & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} \end{bmatrix} \begin{pmatrix} X - X_1 & 0 \\ 0 & X - X_1 \end{pmatrix}^{-1} \\
&\equiv \begin{pmatrix} \mathcal{P}_1 & \mathcal{P}_2 \\ \mathcal{P}_3 & \mathcal{P}_4 \end{pmatrix}
\end{aligned}$$

with

$$\mathcal{P}_1 = Q \begin{bmatrix} J_\infty \left(J_\infty - \frac{a-b}{a+b}J_\infty^{-1} \right)^{-1} & 0 \\ 0 & 0 \end{bmatrix} Q^{-1},$$

$$\mathcal{P}_2 = Q \begin{bmatrix} -\frac{a-b}{a+b} \left(J_\infty - \frac{a-b}{a+b}J_\infty^{-1} \right)^{-1} & 0 \\ 0 & 0 \end{bmatrix} Q^{-1},$$

[2] This is due to the properties of the mapping ψ.

$$\mathcal{P}_3 = Q \begin{bmatrix} \left(J_\infty - \frac{a-b}{a+b} J_\infty^{-1}\right)^{-1} & 0 \\ 0 & 0 \end{bmatrix} Q^{-1},$$

$$\mathcal{P}_4 = Q \begin{bmatrix} -\frac{a-b}{a+b} J_\infty^{-1} \left(J_\infty - \frac{a-b}{a+b} J_\infty^{-1}\right)^{-1} & 0 \\ 0 & 0 \end{bmatrix} Q^{-1},$$

and we have $\mathcal{P}_1 + \mathcal{P}_4 = Q\left(\begin{smallmatrix} I_k & 0 \\ 0 & 0 \end{smallmatrix}\right) Q^{-1} = P$. □

4. Numerical experiments. In this section we illustrate the numerical behavior of the parabolic and elliptic spectral dichotomy. Before considering our test examples, let us recall some qualitative aspects of the proposed methods. When the dichotomy with respect to the imaginary axis is used on the matrix $\mathcal{A}$, the quantity

$$\kappa = 2||\mathcal{A}||\ ||H_{\mathcal{A}}||, \quad \text{where} \quad H_{\mathcal{A}} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} (\mathcal{A} - i\xi)^{-1} (\mathcal{A} - i\xi)^{-*} d\xi, \tag{32}$$

will be referred to as the dichotomy condition number. It was shown in [4] that

$$\tilde{\alpha} \equiv \sup_{\Re\lambda=0} \|(\lambda I_{2n} - \mathcal{A})^{-1}\| \leq \frac{\pi}{3 - 4\log 2}\ \|H_{\mathcal{A}}\| \tag{33}$$

and that the parameter κ can be considered as an indicator of the absence of eigenvalues of the matrix $\mathcal{A}$ on the imaginary axis and within a small neighborhood of it.

Similarly, when the dichotomy with respect to the unit circle is used on the pencil $\lambda\mathcal{B} - \mathcal{A}$, the quantity

$$\omega = ||\mathcal{H}||, \quad \text{where} \quad \mathcal{H} = \frac{1}{2\pi} \int_0^{2\pi} (\mathcal{A} - e^{i\phi}\mathcal{B})^{-1} (\mathcal{A}\mathcal{A}^* + \mathcal{B}\mathcal{B}^*) (\mathcal{A} - e^{i\phi}\mathcal{B})^{-*} d\phi, \tag{34}$$

satisfies [20]

$$\tilde{\beta} \equiv \sup_{|\mu|=1} \|(\mu\mathcal{B} - \mathcal{A})^{-1}\| \leq \max\left\{ \min(||\mathcal{A}||, ||\mathcal{B}||)\ \frac{\pi\omega}{3 - 4\log 2},\ \sqrt{\frac{12}{5}}\omega \right\} \tag{35}$$

and will play the role of κ. That is, if ω is large, then we conclude that the pencil $\lambda\mathcal{B} - \mathcal{A}$ has eigenvalues on a small neighborhood of the unit circle.

For each test example, we plot a figure showing the evolution of the condition number κ (ω) when the parabola (ellipse) varies. We also give in each figure the trace of the projector P that indicates the number of eigenvalues of A outside the parabola (ellipse) and its norm which gives an indication about the angle between the invariant subspaces associated with the eigenvalues inside and outside the parabola (ellipse). The larger the norm of the projector is, the smaller the angle between the two invariant subspaces is.

Example 1. This example is artificially built in an attempt to illustrate the conditions under which the algorithm works better. The matrix under consideration is of order 40 and is of the form

$$A = Q \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} Q^*. \tag{36}$$

The matrix Q is a unitary matrix generated from the QR factorization of a matrix whose elements are chosen randomly from $[-1.0, 1.0]$. The matrix A_{11} is a 20×20

upper triangular matrix whose diagonal elements are on the parabola γ of equation $y^2 = 1-2x$. The other elements are generated randomly between 0 and 1. The matrix A_{22} is also a 20×20 upper triangular matrix whose diagonal elements are such that $A_{22}(k,k) = 1 + \frac{k}{40},\ k = 1, \ldots, 20$ and the other elements are generated randomly between 0 and 1. The matrix A_{12} is a 20×20 random matrix with elements between 0 and 1. The norm of A is 13.41. Its condition number is equal to 1.73×10^3. The condition number of the matrix X of eigenvectors of A is equal to 1.61×10^{15}. The spectrum of A is plotted in Figure 1.

We consider the family of parabolae $y^2 = 2p(p/2 - x)$, where p is a parameter to be varied. Figure 2 shows the evolution of the condition number κ, whose expression is given in (32) when p varies between 0.05 and 6. There are three main branches separated by asymptotes. The regions between asymptotes corresponding to $0.96 \leq p \leq 1.04$ and $1.8 \leq p \leq 3.65$ are such that the parabola $y^2 = 2p(p/2 - x)$ crosses frequently (or is situated in a neighborhood of) some eigenvalues of A. This means that the curve changes very frequently in these regions. In the branch corresponding to $0.05 \leq p \leq 0.96$ we have $P = I_n$; that is, all the eigenvalues are outside the parabola. In the branch corresponding to $1.04 \leq p \leq 1.80$ we have $\mathrm{trace}(P) = 20$ and $||P|| = 3.25 \times 10^4$. In the branch corresponding to $3.65 \leq p \leq 6$ we have $P \equiv 0$; that is, all the eigenvalues are inside the parabola.

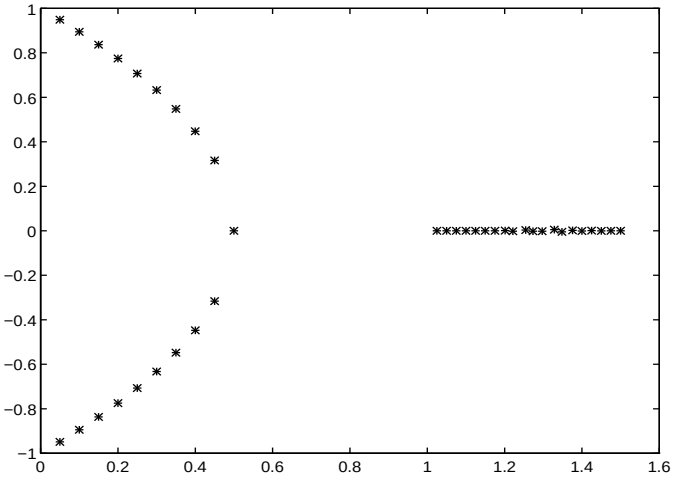


FIG. 1. *Eigenvalue distribution of the matrix in Example* 1.

Table 1 shows approximations of κ and $\tilde{\alpha}$ for different values of the parameter p. The approximation of $\tilde{\alpha}$ is obtained by applying a nonlinear minimization algorithm to the function $x \in \mathbb{R} \to \|(ixI_{2n} - \mathcal{A})^{-1}\|$. The results shown in Table 1 reveal that the parameter $\tilde{\alpha}$ is roughly of the same order of magnitude as $\sqrt{\kappa}$. Actually, it was proven in [2] that $\tilde{\beta}$ can be "as small as" $\sqrt{\omega}$. Note, however, that the computation of κ and ω, or their square root, is much cheaper than that of $\tilde{\alpha}$ and $\tilde{\beta}$. These last parameters may be approximated by either nonlinear minimization algorithms or the bisection algorithm described in [6].

We mentioned in the introduction the possibility of using the Schur-based methods. This may be done by computing the Schur decomposition $Q^*AQ = T = \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix}$, where T_{11} is $m \times m$ and T_{22} is $(n-m) \times (n-m)$. The eigenvalues of T_{11} (T_{22}) are outside (inside) the parabola γ. The spectral projector is then formed

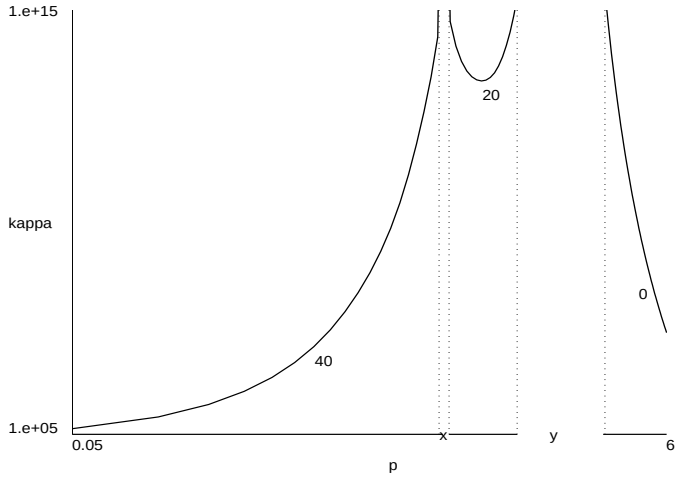


FIG. 2. κ *vs. p.* $x \in [0.96, 1.04]$; $y \in [1.80, 3.65]$.

TABLE 1

p	κ	$\tilde{\alpha}$
0.05	1.35e+05	2.09e+2
0.5	2.16e+08	2.99e+3
0.8	6.78e+11	4.00e+5
1.06	3.73e+14	2.36e+7
1.5	3.26e+13	4.47e+6
2	2.25e+19	7.87e+9
4	1.12e+13	1.34e+6
5	2.0584e+09	3.19e+4

afterwards from the corresponding Schur vectors of $Q = [Q_1\ Q_2]$.

The main problem with this approach is that in finite precision arithmetic the eigenvalues of T are perturbed eigenvalues of A. Hence we are faced with the problem of deciding whether these perturbed eigenvalues are close or not to the contour. One possibility is to compute the ϵ-pseudospectrum of the matrix T [24], but this may be too costly.

Another possibility is to use the quantity $\Delta = 1/\text{sep}(T_{11}, T_{22})$, where $\text{sep}(T_{11}, T_{22})$ denotes the separation between T_{11} and T_{22} [23, 15]. This quantity may be used to indicate the sensitivity of the computed invariant subspace. The advantage of this approach is its moderate complexity compared to the complexity of the dichotomy methods. Indeed, the cost of the Schur-based methods is equal to the cost of the Schur decomposition plus $O(m^3(n-m)^3)$ flops,[3] whereas the cost of the spectral dichotomy lies between $100n^3$ and $200n^3$ [22].

Applied to Example 1, this Schur-based method gives the results shown in Table 2.

Example 2. In this example, taken from [12], we treat only the elliptic spectral

[3] Reliable sep estimates, costing only $O(m^2(n-m)^2 + m(n-m)^2)$ flops, can be obtained by solving triangular Sylvester equations [15].

TABLE 2

p	Δ
$0.05 \le p \le 0.95$	$1.50\mathrm{e}+02$
≈ 1	$4.87\mathrm{e}+15$
$1.05 \le p \le 2.05$	$2.51\mathrm{e}+05$
$2.1 \le p \le 3$	$7.50\mathrm{e}+10 \le \Delta \le 1.00\mathrm{e}+14$
$3.05 \le p \le 6$	$1.50\mathrm{e}+2$

dichotomy. We consider the matrix pencil $\lambda B - A$ of order 20 defined by

$$A = \begin{cases} a_{2i-1,2i-1} = \frac{1}{4} & \text{if } 1 \le i \le 10, \\ a_{2i,2i+2} = 1 & \text{if } 1 \le i \le 9, \\ a_{2i+2,2i} = -5 & \text{if } 1 \le i \le 9, \\ a_{2i,2i-1} = -2 & \text{if } 1 \le i \le 9, \\ a_{i,j} = 0 & \text{otherwise} \end{cases} \quad \text{and } B = \begin{cases} b_{2i,2i} = -\frac{1}{3} & \text{if } 1 \le i \le 10, \\ b_{2i-1,2i+1} = 1 & \text{if } 1 \le i \le 9, \\ b_{2i+1,2i-1} = \frac{1}{5} & \text{if } 1 \le i \le 9, \\ b_{2i,2i+1} = 3 & \text{if } 1 \le i \le 9, \\ b_{i,j} = 0 & \text{otherwise.} \end{cases}$$

The spectrum of this pencil is plotted in Figure 3. We consider a family of ellipses whose major and minor semiaxes are given, respectively, by $a = 3t$ and $b = \frac{1}{t}$; t is a parameter to be varied. Figure 4 shows the evolution of ω when t varies between 0.1 and 1. Starting from the left, we have in the successive branches $\mathrm{trace}(P) = 12$ and $||P|| = 1.09\mathrm{e}+5$; $\mathrm{trace}(P) = 12$ and $||P|| = 1.01\mathrm{e}+5$; $\mathrm{trace}(P) = 10$ and $||P|| = 1.98\mathrm{e}+5$; $\mathrm{trace}(P) = 12$ and $||P|| = 2.03\mathrm{e}+5$; $\mathrm{trace}(P) = 10$ and $||P|| = 6.89\mathrm{e}+5$; $\mathrm{trace}(P) = 12$ and $||P|| = 8.94\mathrm{e}+4$; and finally $\mathrm{trace}(P) = 10$ and $||P|| = 4.66\mathrm{e}+4$. As in the first example, we did not plot the curve situated between the asymptotes since it changes very frequently in these regions.

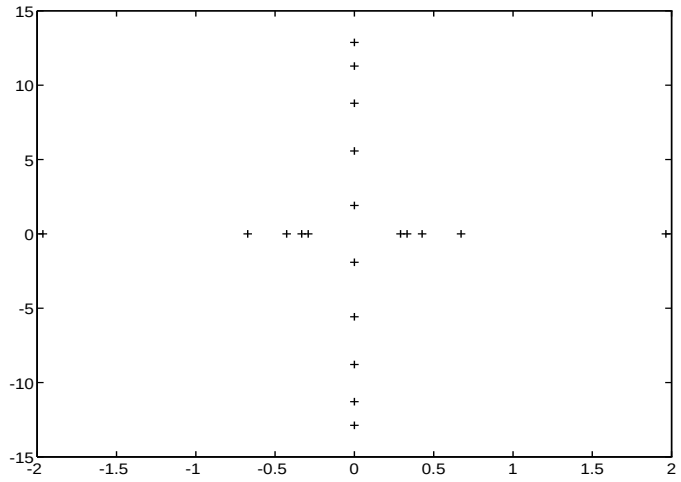


FIG. 3. *Eigenvalue distribution of the matrix pencil in Example* 2.

5. Conclusion. We have proposed some transformation techniques for the spectrum dichotomy problem of a matrix (matrix pencil) with respect to a parabola (ellipse) which use and generalize the cases of the circle and the straight line. Both the theoretical aspect and the numerical treatment of the proposed techniques are illustrated. The proposed spectral transformations can be useful for preparing the initial problems for the matrix sign function methods and projection methods.

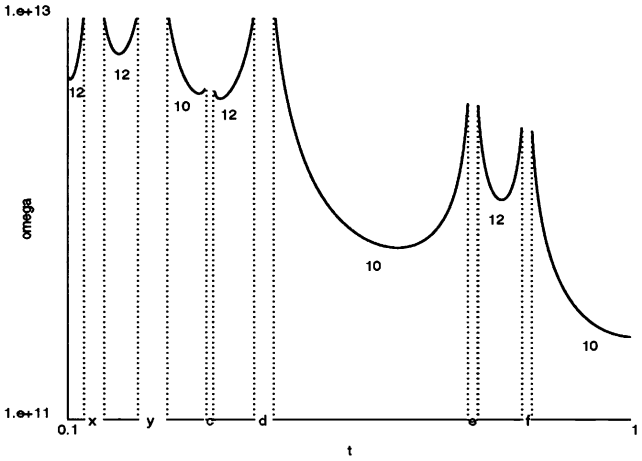


FIG. 4. ω *vs. t.* $a \in [0.107, 0.116]$; $b \in [0.133, 0.150]$; $c \in [0.176, 0.181]$; $d \in [0.214, 0.231]$; $e \in [0.514, 0.535]$; $f \in [0.641, 0.668]$.

It is known that even if the eigenvalues are globally ill conditioned, a subset of them in a given region of the complex plane may be well conditioned. The nice feature of the dichotomy techniques is that the spectral projector associated with the eigenvalues inside and outside the chosen domain, together with an indication of the confidence to be placed in the accuracy of the computed projector, is provided.

Acknowledgments. The authors would like to thank Prof. Bo Kågström and the referees for many useful suggestions.

REFERENCES

[1] L. V. AHLFORS, *Complex Analysis. An Introduction to the Theory of Analytic Functions of One Complex Variable*, McGraw–Hill, New York, 1966.

[2] Z. BAI, J. DEMMEL, AND M. GU, *Inverse Free Parallel Spectral Divide and Conquer Algorithms for Nonsymmetric Eigenproblems*, Tech. rep., Computer Science Division, UCB/CSD-94-793, University of California at Berkeley, 1994.

[3] P. BENNER AND R. BYERS, *An inverse free spectral divide and conquer method for the numerical solution of algebraic Riccati equations*, in IMA Conference on Linear Algebra and Its Applications, University of Manchester, Manchester, UK, 1995.

[4] A. Y. BULGAKOV, *An estimate of the green matrix and the continuity of the dichotomy parameter*, Siberian Math. J., 30 (1989), pp. 139–142.

[5] A. Y. BULGAKOV AND S. K. GODUNOV, *Circular dichotomy of the spectrum of a matrix*, Siberian Math. J., 29 (1988), pp. 734–744.

[6] R. BYERS, *A bisection method for measuring the distance of a stable matrix to the unstable matrices*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 875–881.

[7] R. BYERS, C. HE, AND V. MEHRMANN, *The Matrix Sign Function Method and the Computation of Invariant Subspaces*, Tech. rep., Fakultät für Mathematik, TU Chemnitz-Zwickau, Chemnitz, FRG, 1994.

[8] J. W. DEMMEL, *Computing stable eigendecomposition of matrices*, Linear Algebra Appl., 79 (1986), pp. 163–193.

[9] J. W. DEMMEL AND B. KÅGSTRÖM, *Computing stable eigendecomposition of matrix pencil*, Linear Algebra Appl., 88-89 (1987), pp. 139–186.

[10] F. R. GANTMACHER, *Théorie des matrices. Tome* 2, Dunod, Paris, 1966.

[11] S. K. GODUNOV, *Problem of the dichotomy of the spectrum of a matrix*, Siberian Math. J., 27 (1986), pp. 649–660.

[12] S. K. Godunov and M. Sadkane, *Elliptic dichotomy of a matrix spectrum*, Linear Algebra Appl., 248 (1996), pp. 205–232.
[13] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[14] B. Kågström and P. Van Dooren, *A generalized state-space approach for the additive decomposition of a transfer matrix*, J. Numer. Linear Algebra Appl., 1 (1992), pp. 165–181.
[15] B. Kågström and P. Poromaa, *Computing eigenspaces with specified eigenvalues of a regular matrix pair* (A, B) *and condition estimation: Theory, algorithm and software*, Numer. Algorithms, to appear.
[16] C. Kenney and G. Hewer, *The sensitivity of the algebraic and differential Riccati equations*, SIAM J. Control Optim., 28 (1990), pp. 50–69.
[17] C. S. Kenney and A. J. Laub, *The matrix sign function*, IEEE Trans. Automat. Control, 40 (1995), pp. 1330–1348.
[18] A. J. Laub, *Invariant subspace methods for the numerical solution of algebraic Riccati equations*, in The Riccati equation, S. Bittanti, A. Laub, and J. Willems, eds., Springer-Verlag, Berlin, 1991, pp. 163–196.
[19] A. N. Malyshev, *Computing invariant subspaces of a regular linear pencil of matrices*, Siberian Math. J., 30 (1989), pp. 559–567.
[20] A. N. Malyshev, *Guaranteed accuracy in spectral problems of linear algebra*, Siberian Adv. Math. J. I, II, 2 (1992), pp. 144–197.
[21] A. N. Malyshev, *Parallel algorithm for solving some spectral problems of linear algebra*, Linear Algebra Appl., 188–189 (1993), pp. 489–520.
[22] A. N. Malyshev, *Parallel Aspects of Some Spectral Problems in Linear Algebra*, Tech. rep. NM-R9113, Department of Numerical Mathematics, CWI, Amsterdam, The Netherlands, 1991.
[23] G. Stewart and J. Sun, *Matrix Perturbation Theory*, Academic Press, San Diego, CA, 1990.
[24] L. N. Trefethen, *Pseudospectra of matrices*, in Numerical Analysis 1991, D. F. Griffiths and G. A. Watson, eds., Longman Scientific and Technical, Harlow, UK, 1992.
[25] C. F. Van Loan, *How near is a stable matrix to an unstable matrix?*, in Contemporary Mathematics, Proc. of the summer research conference, vol. 47, AMS, Providence, RI, 1985, pp. 465–478.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, pp. 279–290, April 1997

002

LEAST-SQUARES APPROXIMATE SOLUTION OF OVERDETERMINED SYLVESTER EQUATIONS*

A. SCOTTEDWARD HODEL† AND PRADEEP MISRA‡

Abstract. We address the problem of computing a low-rank estimate Y of the solution X of the Lyapunov equation $AX + XA' + Q = 0$ *without* computing the matrix X itself. This problem has applications in both the reduced-order modeling and the control of large dimensional systems as well as in a hybrid algorithm for the rapid numerical solution of the Lyapunov equation via the alternating direction implicit method. While no known methods for low-rank approximate solution provide the two-norm optimal rank k estimate X_k of the exact solution X of the Lyapunov equation, our iterative algorithms provide an effective method for estimating the matrix X_k by minimizing the error $\|AY + YA' + Q\|_F$.

Key words. Sylvester equation, least squares, iterative, conjugate gradient

AMS subject classifications. 15A06, A5A24, 65F05

PII. S0895479893252337

1. Introduction.

The Lyapunov equation

$$AX + XA' + Q = 0, \tag{1.1}$$

$A, Q \in \mathbb{R}^{n\times n}, Q = Q'$ plays a significant role in numerous problems in control, communication systems theory, and power systems. Recent applications of the Lyapunov equation include the design of reduced-order state estimators and controllers [2], [20], [28], [29] and the solution of robust decentralized control problems [30], [33]. The Lyapunov equation also has applications in stability analysis [21], [25]. Standard methods for the numerical solution of the Lyapunov equation [1], [12] make use of the real Schur decomposition $A = USU'$, where U is an orthogonal matrix and S is quasi-upper triangular. The matrix U is used to transform the Lyapunov equation (1.1) into a form that is readily solved through forward substitution. More recently, Lu [26] and Wachspress [34] proposed the use of the alternating direction implicit (ADI) method for the iterative solution of Lyapunov equations for which all eigenvalues of the matrix A (or of $-A$) are in the right half of the complex plane.

Recently proposed numerical techniques for the numerical solution of the Lyapunov equation have involved iterative solution techniques [23], [19], [31] or low-rank approximate solution techniques [17], [18], [22]. Each of these methods requires the numerical solution of either a reduced-order Lyapunov equation

$$(V'AV)\Sigma_V + \Sigma_V(V'A'V') + V'QV = 0 \tag{1.2}$$

or a least-squares problem

$$\Sigma_V = \arg\min \|AV\Sigma_V V' + V\Sigma_V V'A' + Q\|_F . \tag{1.3}$$

*Received by the editors July 19, 1993; accepted for publication (in revised form) by S. J. Hammarling April 1, 1996. This work was supported in part by NSF grant ECS-9110083.
http://www.siam.org/journals/simax/18-2/25233.html
†Department of Electrical Engineering, 200 Broun Hall, Auburn University, Auburn, AL 36849 (scotte@eng.auburn.edu).
‡RC-311, Department of Electrical Engineering, Wright State University, Dayton, OH 45435 (pmisra@valhalla.cs.wright.edu).

The numerical solution of generalized Lyapunov equations

$$AXB' + BXA' + C = 0$$

may be achieved through the use of a QZ decomposition [27] of the matrix pencil (A, B) [7], [8]; low-rank approximate solution techniques may be applied to these problems in a fashion analogous to the standard case (1.1).

In this paper, we address the least-squares solution of minimizations of the form

$$\min_X \|AXB' + CXD' + F\|_F, \tag{1.4}$$

where (for simplicity in exposition) A, B, C, and $D \in \mathbb{R}^{n\times k}$, $X \in \mathbb{R}^{k\times k}$, $F \in \mathbb{R}^{n\times n}$, and $k \ll n$. Note that (1.3) then simply reduces to a special case of (1.4). The minimization (1.4) can be transformed to a minimization of the form

$$\min \left\|\bar{A}\bar{x} + \bar{b}\right\|_2 \tag{1.5}$$

through a Kronecker product expansion; see [24]. Techniques for the solution of large, sparse least-squares problems (1.5) have been addressed in several iterative algorithms, e.g., [9], [13], [32]. It should be noted that, unlike the Kronecker product expansion of the Lyapunov equation (1.1), the Kronecker product expansion (1.5) of the least-squares minimization (1.4) yields a *dense* matrix $\bar{A}$ in general, since no sparsity structure can be assumed for the matrices A, B, C, D, and F in applications that do not involve Krylov subspaces [15], [16], [18], [19].

It should be noted that a difficulty associated with flexible structures (second-order PDEs) that does not usually occur in heat flow problems is that the discretizations $\dot{x} = Ax + Bu$ do not automatically satisfy the constraint $A + A' < 0$ discussed in [15] and [16]. Hence, a least-squares approach as proposed in this paper becomes preferable to a reduced-order Lyapunov equation (the approach studied at length in [15] and [31]).

In the case $k = n$, (1.4) becomes a generalized Sylvester equation

$$AXB' + CXD' + F = 0, \tag{1.6}$$

which can be solved by reduction of the matrix pencils (A, C) and (D, B) to Schur-triangular form and Hessenberg-triangular form [27], respectively, and then by applying a modified version of the Golub–Nash–Van Loan algorithm [10]. Unfortunately, this approach is not directly applicable to the minimization (1.4); in particular, if $\operatorname{rank}\left(\begin{bmatrix} A & C \end{bmatrix}\right) = 2k$ then all of the generalized eigenvalues of the pencil $(A - \lambda C)$ are zero, and no useful decomposition of the problem can be obtained. However, the solution of (1.6) plays a key role in our algorithm for the solution of (1.4).

We propose the numerical solution of the minimization (1.4) through a preconditioned conjugate gradient (CG) algorithm [6]; the development of our algorithm is as follows. First, in section 2 we present an overview of Krylov subspace techniques as related to the numerical solution of the Lyapunov equation. In section 3 we give an overview of the minimization of (1.4) and present algorithms for its numerical solution in section 4. Following this, we present numerical examples in section 5. In section 6 we make some concluding remarks.

2. Krylov subspaces and iterative techniques. Krylov subspace techniques have gained increasing popularity in the solution of large, sparse systems of linear equations

$$Ax = b. \tag{2.1}$$

A Krylov subspace $\mathcal{K}(A, v, k)$ is defined as

$$\mathcal{K}(A, v, k) = \text{span}(\begin{bmatrix} v & Av & \cdots & A^{k-1}v \end{bmatrix}),$$

where $A \in \mathbb{R}^{n\times n}$, $v \in \mathbb{R}^n$, and k is an integer. One typically uses the Arnoldi algorithm [11] or a variation thereof to compute an orthogonal matrix $V_k \in \mathbb{R}^{n\times k}$, or simply V, such that $\text{span}(V) = \mathcal{K}(A, v, k)$. The Arnoldi algorithm generates a sequence of orthogonal matrices V_k such that $AV_k = V_{k+1}H_{k+1}$, where $H_{k+1} \in \mathbb{R}^{k+1\times k}$ is an upper Hessenberg matrix; i.e., $i > j + 1 \Rightarrow H_{ij} = 0$.

The GMRES algorithm [32] uses Krylov subspace bases V_k obtained by the Arnoldi algorithm to "project" the underlying problem (2.1) into a low-rank minimization

$$y^* = \arg\min_y \left\| H_{k+1}y - {V_{k+1}}'b \right\|_2$$

and approximates the solution x of (2.1) as $x \approx V_k y^*$. If the corresponding residual is too large, then the algorithm may either (1) increase the dimension k of the Krylov subspace or (2) use iterative refinement on the residual with a rank k Krylov subspace (GMRES(k)). Barring algorithm stagnation due to the identification of an A-invariant subspace (a consequence of catastrophic breakdown in the Arnoldi method) [4], the iterative application of GMRES(k) guarantees monotone decreasing residuals corresponding to each iteration.

Hu and Reichel [19] propose an iterative algorithm, based on GMRES [32], for the solution of large, sparse Sylvester equations

$$AX + XB + C = 0. \tag{2.2}$$

Each iteration of the Hu–Reichel algorithm uses Krylov subspaces of G and H to construct a minimization (1.4) with $\text{rank}\left(\begin{bmatrix} A & C \end{bmatrix}\right) = \text{rank}\left(\begin{bmatrix} B & D \end{bmatrix}\right) = k + 1$ whose solution X is obtained by a CG algorithm. A related approach is proposed by Jaimoukha and Kasenally [23].

Hodel and Poolla [17] and Hodel, Tenison, and Poolla [18] iteratively compute estimates of the dominant invariant subspace of the solution X of the Lyapunov equation (1.1). Similarly, Hodel [16] proposes gradient-based schemes that attempt to identify a low-rank subspace basis V that minimizes the associated residual of the Lyapunov equation. Since each of these algorithms identifies a subspace basis $V \in \mathbb{R}^{n\times k}$ and not a low-rank approximate solution $\hat{X} \in \mathbb{R}^{n\times n}$ of the Lyapunov equation (1.1), either of these algorithms may be used in tandem with the minimization of (1.4) to obtain a low-rank estimate $\hat{X} = V\Sigma V'$, where Σ is computed from (1.4). This approach does not necessarily yield estimates $\hat{X}$ that lie in a Krylov subspace.

Saad [31] obtains a low-rank approximate solution of the Lyapunov equation (1.1) by applying Krylov subspaces to the identity

$$X = \int_0^\infty e^{A't} Q e^{At} dt, \tag{2.3}$$

where A is stable (all eigenvalues lie in the left half plane); the evaluation of this integral is clearly undesirable when A is not stable. This algorithm computes an estimate $\hat{X} = V\Sigma V'$ of X by solving a reduced-order Lyapunov equation (1.2). This approach is applied in [22] to construct low-order models/controllers for very large, sparse linear dynamic systems. While error bounds are available for this approach, care must be taken in its application, especially when the matrix $(A + A')$ is not negative definite; see [17]. Further issues in the use of the integral (2.3) are discussed in [15].

3. Reduction of problem dimension. The minimization (1.4) can be rewritten as a standard least-squares problem (1.5) through a Kronecker product expansion. More precisely, in (1.5) we let $\bar{A} = (B \otimes A + D \otimes C)$, $\bar{x} = \mathrm{vec}(X)$, and $\bar{b} = \mathrm{vec}(F)$, where $Y \otimes Z = \left[\ y_{ij}Z\ \right]$ is the Kronecker product of two arbitrary matrices Y and Z, and $\mathrm{vec}(A)$ is the vector stack of the matrix A; e.g., if $Z \in \mathbb{R}^{n\times m}$, then

$$\mathrm{vec}(Z) = \left[\ Z_{\cdot 1}{}' \quad \cdots \quad Z_{\cdot m}{}'\ \right]',$$

where $Z_{\cdot j}$ is the jth column of the matrix Z. (Observe that $\mathrm{vec}(ABC) = (C' \otimes A)\,\mathrm{vec}(B)$ [3].) The Kronecker product expansion of (1.4) yields an overdetermined sparse system of n^2 equations in k^2 unknowns so that a naive application of a QR algorithm would require $O(n^2k^4 + k^6)$ flops to obtain the optimal solution X. If $k < n/3$, then the dimension of the minimization may be reduced, as shown in the following lemma.

LEMMA 3.1. *Let $A, B, C, D \in \mathbb{R}^{n\times k}, F \in \mathbb{R}^{n\times n}$, and let $A_1, A_2, B_1, C_1, D_1, D_2 \in \mathbb{R}^{k\times k}$ satisfy the QR factorizations*

$$\left[\ Q_1^{(1)} \quad Q_2^{(1)} \quad Q_3^{(1)}\ \right] \begin{bmatrix} C_1 & A_1 \\ 0 & A_2 \\ 0 & 0 \end{bmatrix} = \left[\ C \quad A\ \right],$$

$$\left[\ Q_1^{(2)} \quad Q_2^{(2)} \quad Q_3^{(2)}\ \right] \begin{bmatrix} B_1 & D_1 \\ 0 & D_2 \\ 0 & 0 \end{bmatrix} = \left[\ B \quad D\ \right],$$

where $Q_1^{(j)}, Q_2^{(j)} \in \mathbb{R}^{n\times k}$ and $Q_3^{(j)} \in \mathbb{R}^{n\times n-2k}$, and $[Q_1^{(j)} \quad Q_2^{(j)} \quad Q_3^{(j)}]$ is an orthogonal basis of $\mathbb{R}^n$, $j = 1, 2$. Then $X \in \mathbb{R}^{k\times k}$ minimizes $\|AXB' + CXD' + F\|_F$ if and only if X minimizes

$$\left\| \begin{bmatrix} A_1XB_1{}' + C_1XD_1{}' \\ A_2XB_1{}' \\ C_1XD_2{}' \end{bmatrix} + \begin{bmatrix} \hat{F}_{11} \\ \hat{F}_{21} \\ \hat{F}_{12} \end{bmatrix} \right\|_F, \tag{3.1}$$

where $\hat{F}_{ij} = Q_i^{(1)}{}' F Q_j^{(2)}$.

Proof. Let $Q_j = [Q_1^{(j)} \quad Q_2^{(j)} \quad Q_3^{(j)}]$, $j = 1, 2$, so that

$$\left[\ C \quad A\ \right] = Q_1 \begin{bmatrix} C_1 & A_1 \\ 0 & A_2 \\ 0 & 0 \end{bmatrix}, \qquad \left[\ B \quad D\ \right] = Q_2 \begin{bmatrix} B_1 & D_1 \\ 0 & D_2 \\ 0 & 0 \end{bmatrix}$$

and define

$$\hat{F} = Q_1{}' F Q_2 = \begin{bmatrix} \hat{F}_{11} & \hat{F}_{12} & \hat{F}_{13} \\ \hat{F}_{21} & \hat{F}_{22} & \hat{F}_{23} \\ \hat{F}_{31} & \hat{F}_{32} & \hat{F}_{33} \end{bmatrix}.$$

Then

$$\begin{aligned} &\min \|AXB' + CXD' + F\|_F \\ &\quad = \min \left\| Q_1{}' \left(AXB' + CXD' + F\right) Q_2 \right\|_F \end{aligned}$$

$$= \min \left\| \begin{bmatrix} A_1 \\ A_2 \\ 0 \end{bmatrix} X \begin{bmatrix} B_1{}' & 0 & 0 \end{bmatrix} + \begin{bmatrix} C_1 \\ 0 \\ 0 \end{bmatrix} X \begin{bmatrix} D_1{}' & D_2{}' & 0 \end{bmatrix} + \hat{F} \right\|_F$$

$$= \min \left\| \begin{bmatrix} A_1 X B_1{}' + C_1 X D1' + \hat{F}_{11} & C_1 X D_2{}' + \hat{F}_{12} & \hat{F}_{13} \\ A_2 X B_1{}' + \hat{F}_{21} & \hat{F}_{22} & \hat{F}_{23} \\ \hat{F}_{31} & \hat{F}_{32} & \hat{F}_{33} \end{bmatrix} \right\|_F .$$

Since $\hat{F}_{13}, \hat{F}_{22}, \hat{F}_{23}, \hat{F}_{31}, \hat{F}_{32}$, and $\hat{F}_{33}$ are constant for all values of X, the above minimization is unaffected by these terms, and the lemma follows. ∎

Remark 3.1. The practical reduction of (1.4) to (3.1) for the general case (F does not possess a sparse or other exploitable structure) can be accomplished in $O(n^2k)$ flops as follows.

1. Compute and store the Householder vectors $h_i^{(j)}$, $i = 1, 2k$, $j = 1, 2$, obtained in the QR factorizations of $\begin{bmatrix} C & A \end{bmatrix}$ and $\begin{bmatrix} D & B \end{bmatrix}$ in $O(nk^2)$ flops.
2. Accumulate Householder reflections $H_i^{(j)} = (I - (2/h_i^{(j)\prime} h_i^{(j)}) h_i^{(j)} h_i^{(j)\prime})$ to obtain
$$\bar{Q}_j = \begin{bmatrix} Q_1^{(j)} & Q_2^{(j)} \end{bmatrix},$$
$i = 1, \ldots, 2k$, $j = 1, 2$, in $O(nk^2)$ flops.
3. Compute $Z = F\bar{Q}_2 \in \mathbb{R}^{n \times 2k}$ in $O(n^2k)$ flops.
4. Compute
$$\begin{bmatrix} \hat{F}_{11} \\ \hat{F}_{21} \end{bmatrix} = \bar{Q}_1' Z \text{ and } \hat{F}_{12} = Q_1^{(1)\prime} Z \begin{bmatrix} 0_{k \times k} \\ I_k \\ 0 \end{bmatrix}$$
(i.e., multiply by the last k columns of Z) in $O(nk^2)$ flops.

Observe that the dominant computational cost of $O(n^2k)$ flops occurs in step 3. This cost can be greatly reduced if matrix-vector products Fv can be computed in much less than n^2 flops, e.g., if F is sparse or low-rank. The latter will be the case in controller/model reduction applications such as [28]. Since the number of inputs/outputs is greatly exceeded by the number of states in a typical dynamic system, the matrix F is given by $\hat{B}\hat{B}'$, $B \in \mathbb{R}^{n \times m}$, $m \ll n$. Then step 3 above can be computed in $O(mnk)$ time, rendering the overall complexity to $O(nmk)$ or $O(nk^2)$ (whichever is smaller).

Remark 3.2. If the least-squares minimization (1.4) is obtained via Krylov subspaces as in [19], then the corresponding minimization has $A, B, C, D \in \mathbb{R}^{(k+1) \times k}$, which obviates the need for the above reduction.

4. Iterative solution by CG methods. We shall henceforth assume that the minimization (1.4) has been posed in the form of Lemma 3.1; i.e., we wish to solve the least-squares problem

$$\min \left\| \begin{bmatrix} A_1 X B_1{}' + C_1 X D_1{}' \\ A_2 X B_1{}' \\ C_1 X D_2{}' \end{bmatrix} + \begin{bmatrix} \hat{F}_{11} \\ \hat{F}_{21} \\ \hat{F}_{12} \end{bmatrix} \right\|_F . \tag{4.1}$$

As in the case of (1.4), (4.1) can be solved by a Kronecker product expansion (1.5) with

$$\bar{A} = \begin{bmatrix} L_1 \\ L_2 \\ L_3 \end{bmatrix} \triangleq \begin{bmatrix} B_1 \otimes A_1 + D_1 \otimes C_1 \\ B_1 \otimes A_2 \\ D_2 \otimes C_1 \end{bmatrix}, \tag{4.2}$$

$$\bar{x} = \mathrm{vec}(X), \text{ and } \bar{b} = \begin{bmatrix} \bar{b}_1 \\ \bar{b}_2 \\ \bar{b}_3 \end{bmatrix} \triangleq \begin{bmatrix} \mathrm{vec}(\hat{F}_{11}) \\ \mathrm{vec}(\hat{F}_{21}) \\ \mathrm{vec}(\hat{F}_{12}) \end{bmatrix}.$$

The QR method allows the computation of a matrix X that minimizes (4.1) in $O(k^6)$ flops.

We may also minimize (4.1) in $O(k^5)$ flops by applying the CG algorithm [14] to the normal equations $\bar{A}'\bar{A}\bar{x} = -\bar{A}'\bar{b}$ as follows.

ALGORITHM 1. Solution of (4.1) by CGs.

Inputs $A_1, A_2, B_1, C_1, D_1, D_2, \hat{F}_{11}, \hat{F}_{21}, \hat{F}_{12} \in \mathbb{R}^{k\times k}$.

Outputs $X \in \mathbb{R}^{k\times k}$ satisfying the minimization (4.1).

1. $X_0 = 0$, $j = 0$, $R_{11}^{(0)} = \hat{F}_{11}$, $R_{21}^{(0)} = \hat{F}_{21}$, $R_{12}^{(0)} = \hat{F}_{12}$,

$$R_0 = A_1{}'\hat{F}_{11}B_1 + C_1{}'\hat{F}_{11}D_1 + A_2{}'\hat{F}_{21}B_1 + C_1{}'\hat{F}_{12}D_2$$

2. while $R_j \neq 0$
   (a) $j = j + 1$
   (b) if $j = 1$
   $P_j = R_0$
   (c) else $\beta_j = \dfrac{\mathrm{vec}(R_{j-1})'\mathrm{vec}(R_{j-1})}{\mathrm{vec}(R_{j-2})'\mathrm{vec}(R_{j-2})}$, $P_j = R_{j-1} + \beta_j P_{j-1}$
   (d) end if
   (e) Compute $W_{11}^{(j)} = A_1P_jB_1{}' + C_1P_jD_1{}'$, $W_{21}^{(j)} = A_2P_jB_1{}'$, and $W_{12}^{(j)} = C_1P_jD_2{}'$, and let $W_j = \begin{bmatrix} \mathrm{vec}(W_{11}^{(j)})' & \mathrm{vec}(W_{21}^{(j)})' & \mathrm{vec}(W_{12}^{(j)})' \end{bmatrix}'$.
   (f) $\alpha_j = \dfrac{\mathrm{vec}(R_{j-1})'\mathrm{vec}(R_{j-1})}{\mathrm{vec}(P_j)'\bar{A}p_j}$
   (g) $X_j = X_{j-1} + \alpha_j P_j$
   (h) Compute residuals

$$R_{11}^{(j)} = A_1X_{j-1}B_1{}' + C_1X_{j-1}D_1{}' + \hat{F}_{11} \qquad R_{21}^{(j)} = A_2X_{j-1}B_1{}' + \hat{F}_{21}$$
$$R_{12}^{(j)} = C_1X_{j-1}D_2{}' + \hat{F}_{12}$$
$$R_j = A_1{}'R_{11}^{(j)}B_1 + C_1{}'R_{11}^{(j)}D_1 + A_2{}'R_{21}^{(j)}B_1 + C_1{}'R_{12}^{(j)}D_2$$

3. end while
4. $X = X_j$

In exact arithmetic, the CG algorithm will converge in at most k^2 iterations; if $\bar{A}'\bar{A}$ is a rank l modification to the identity matrix, then the CG algorithm will converge in at most l iterations; see [11] and [14] for details. The chief disadvantage of the CG method is the loss of $(\bar{A}'\bar{A})$ orthogonality between the vectors p_j as j increases; that is, computed vectors p_j do not satisfy the relation $p_j{}'\bar{A}'\bar{A}\begin{bmatrix} p_1 & \cdots & p_{j-1} \end{bmatrix} = 0$. Because of this numerical behavior, the CG algorithm has come to be regarded as a purely iterative method for large, sparse linear systems of equations. However, if rank $\left(L_2{}'L_2 + L_3{}'L_3\right)$ is not too large, as in [19], then convergence can be accelerated by using a preconditioned conjugate gradient (PCG) algorithm [6]. The PCG algorithm is based on the use of a splitting $(\bar{A}'\bar{A}) = M + N$, where M is symmetric, positive definite, and easy to invert and "near" $\bar{A}$. The PCG algorithm is as follows.

ALGORITHM 2. Generalized CGs.

Inputs $M, N \in \mathbb{R}^{n\times n}$, both symmetric, M positive definite and (by assumption) $\bar{A} = M + N$ positive definite, and $b \in \mathbb{R}^n$.

Outputs $x \in \mathbb{R}^n$ satisfying $\bar{A}x + b = 0$.

1. $x_{-1} = x_0 = 0$, $j = 0$, $r_0 = b$
2. while $r_j \neq 0$
   (a) $j = j + 1$; compute $z_j := -M^{-1} r_{j-1}$.
   (b) $\gamma_j = \dfrac{z_j{}' M z_j}{z_j{}' \bar{A} z_j}$
   (c) $\omega_j = 1$ if $j = 1$, else $\omega_j = \left(1 - \dfrac{\gamma_j (z_j{}' M z_j)}{\omega_{j-1}\gamma_{j-1}(z_{j-1}{}' M z_{j-1})}\right)^{-1}$
   (d) $x_j = x_{j-2} + \omega_j \left(\gamma_j z_j + x_{j-1} - x_{j-2}\right)$.
   (e) $r_j = \bar{A}x_j + \bar{b}$.
3. end while
4. $x = x_j$

We apply the PCG algorithm to the minimization (4.1) as follows. The normal equations corresponding to (4.1) are

$$(L'L)\, x = -L'\bar{f}.$$

Observe that $L'L = L_1{}'L_1 + L_2{}'L_2 + L_3{}'L_3$, and so this problem decomposes naturally to the splitting $M = L_1{}'L_1$, $N = \left(L_2{}'L_2 + L_3{}'L_3\right)$. In order to solve $Mz_j = -r_j$, observe that if L_1 is nonsingular then z_j satisfies

$$\begin{aligned} 0 &= L_1{}'L_1 z_j + L_1{}' r_1^{(j-1)} + L_2{}' r_2^{(j-1)} + L_3{}' r_3^{(j-1)} \\ &= L_1 z_j + r_1^{(j-1)} + L_1{}^{-T}\left(L_2{}' r_2^{(j-1)} + L_3{}' r_3^{(j-1)}\right), \end{aligned}$$

which may be solved in matrix form in $O(k^3)$ flops as

$$\begin{aligned} 0 &= A_1{}' T_j B_1 + C_1{}' T_j D_1 - \left(A_2{}' R_{21}^{(j-1)} B_1 + C_1{}' R_{12}^{(j-1)} D_2\right), \\ 0 &= A_1 Z_j B_1{}' + C_1 Z_j D_1{}' + (R_{11}^{(j-1)} + T_j). \end{aligned}$$

The resulting matrix-valued PCG algorithm is shown below.

ALGORITHM 3. Solution of (4.1) by PCGs.

Inputs $A_1, A_2, B_1, C_1, D_1, D_2, \hat{F}_{11}, \hat{F}_{21}, \hat{F}_{12} \in \mathbb{R}^{k\times k}$.

Outputs $X \in \mathbb{R}^{k\times k}$ satisfying the minimization (4.1).

1. $X_{-1} = X_0 = 0$, $j = 0$, $R_{11}^{(0)} = \hat{F}_{11}$, $R_{21}^{(0)} = \hat{F}_{21}$, $R_{12}^{(0)} = \hat{F}_{12}$,

$$R_0 = A_1{}'\hat{F}_{11}B_1 + C_1{}'\hat{F}_{11}D_1 + A_2{}'\hat{F}_{21}B_1 + C_1{}'\hat{F}_{12}D_2$$

2. while $R_j \neq 0$
   (a) $j = j + 1$; Solve for T_j and Z_j:

$$\begin{aligned} 0 &= A_1{}' T_j B_1 + C_1{}' T_j D_1 - \left(A_2{}' R_{21}^{(j-1)} B_1 + C_1{}' R_{12}^{(j-1)} D_2\right) \\ 0 &= A_1 Z_j B_1{}' + C_1 Z_j D_1{}' + (R_{11}^{(j-1)} + T_j) \end{aligned}$$

   (b) Compute $W_{11}^{(j)} = A_1 Z_j B_1{}' + C_1 Z_j D_1{}'$, $W_{21}^{(j)} = A_2 Z_j B_1{}'$, and $W_{12}^{(j)} = C_1 Z_j D_2{}'$, and let $W_j = \left[\ \mathrm{vec}(W_{11}^{(j)})'\quad \mathrm{vec}(W_{21}^{(j)})'\quad \mathrm{vec}(W_{12}^{(j)})'\ \right]'$.
   (c) $\gamma_j = \dfrac{-\mathrm{vec}(Z_j)'\mathrm{vec}(R_{j-1})}{\mathrm{vec}(W_j)'\mathrm{vec}(W_j)}$

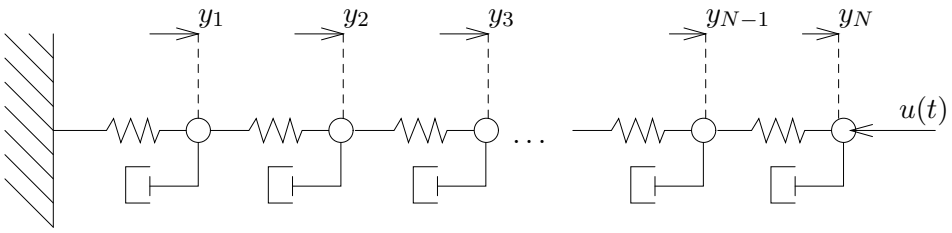


FIG. 5.1. *Mass-spring-dashpot system.*

(d) $\omega_j = 1$ if $j = 1$, else $\omega_j = \left(1 - \dfrac{\gamma_j (z_j{}' r_{j-1})}{\omega_{j-1}\gamma_{j-1}(z_{j-1}{}' r_{j-2})}\right)^{-1}$

(e) $X_j = X_{j-2} + \omega_j\,(\gamma_j Z_j + X_{j-1} - X_{j-2})$.

(f) Compute residuals

$$R_{11}^{(j)} = A_1 X_{j-1} B_1{}' + C_1 X_{j-1} D_1{}' + \hat{F}_{11} \qquad R_{21}^{(j)} = A_2 X_{j-1} B_1{}' + \hat{F}_{21}$$
$$R_{12}^{(j)} = C_1 X_{j-1} D_2{}' + \hat{F}_{12}$$
$$R_j = A_1{}' R_{11}^{(j)} B_1 + C_1{}' R_{11}^{(j)} D_1 + A_2{}' R_{21}^{(j)} B_1 + C_1{}' R_{12}^{(j)} D_2$$

3. end while
4. $X = X_j$

Remark 4.1. If rank $\left(\begin{bmatrix} L_2 & L_3 \end{bmatrix}\right)$ is small ($\ll k^2$), then in exact arithmetic this algorithm should converge much faster than CG (Algorithm 1); see [6, pp. 319–320]. For example, minimizations (4.1) that arise in [19] can be solved in only $2k$ iterations, or $O(k^4)$ work. However, it should be pointed out that the PCG algorithm is not necessarily numerically superior to the CG algorithm; in particular, the operator M is explicitly inverted in step 2a of Algorithm 1; this is undesirable when L_1 is poorly conditioned.

Remark 4.2. Unfortunately, it is not immediately obvious how the conditioning of L_1 relates to the original matrices A, B, C, D. Hence, an explicit bound on the conditioning of $L_1{}' L_1$ appears impossible to determine. However, one may use Byers's condition estimator [5] to determine when an ill-conditioned system occurs; a variant of this algorithm may be employed with the preconditioner in the present algorithm. The response to an ill-conditioned estimator depends on the scenario in which it occurs; one may simply increase the dimension of V (as in Krylov subspace-based algorithms) or one may dispense with the preconditioner to use either the CG algorithm or (if applicable) the algorithms presented in [19] or [23]. When well conditioned, the PCG algorithm in this paper provides an improvement in algorithm speed.

5. Numerical examples. Algorithms 1 and 3 were tested on a lumped mass-spring-damper model of a vibrating system (see Figure 5.1); such models arise in numerous engineering applications. In Figure 5.1, y_j denotes the displacement of mass j from its rest position; $u(t)$ is an external (controlled) force; and all N masses, springs, and dashpots are assumed to be identical with mass m, stiffness ρ, and damping δ, respectively. The first-order dynamic model of the system is

$$\dot{x} = Ax + Bu,$$

TABLE 5.1
Flop counts vs. k for implementations of Algorithms 1 *and* 3.

k	Algorithm 1	Algorithm 3
5	6635	3.918E+04
10	5.043E+04	3.232E+05
15	1.674E+05	1.172E+06

where $A = \left[\begin{smallmatrix} 0 & I_n \\ A_{21} & -(\delta/m)I_n \end{smallmatrix}\right] \in \mathbb{R}^{2N\times 2N}$,

$$A_{21} = \frac{\rho}{m}\begin{bmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & -2 & 1 \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix}, \text{ and } B = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

Example systems were run with $N = 100$, 200, 300, and 400. Results presented in this section are for $N = 300$ with parameters (ρ, δ, m) selected as either $(1, 0.1, 1)$ or $(10, 10^{-3}, 10^{-2})$, respectively. The second set of parameters yields a very lightly damped system.

A solution X was sought to the minimization

$$\min_{X\in\mathbb{R}^{k\times k}} \|(AV)XV' + VX(A'V') + BB'\|_F,$$

where V was an orthogonal basis of the Krylov subspace

$$\text{span}\left(\begin{bmatrix} b & Ab & \cdots & A^{k-1}b \end{bmatrix}\right)$$

for $k = 5, 10$, and 15. Numerical implementation of Algorithms 1 and 3 was done using MATLAB version 4.2a on a Sun Sparc-10. Table 5.1 shows returned flop counts per iteration for the two algorithms vs. problem dimension parameter k. Figure 5.2 shows the plots of the residual of the normal equations for the CG and PCG iterations; system parameters were $\delta = 0.1$, $\rho = 1$, and $m = 1$. Observe that the PCG method residual reaches its equilibrium value in roughly k iterations, consistent with its expected convergence behavior. Both algorithms are sensitive to the condition of the underlying system; Figure 5.3 shows the residuals for $\delta = 10^{-3}$, $\rho = 10$, and $m = 0.01$. The deterioration in performance is due to the wide spread in singular values of $L'L$ associated with lightly damped, high-frequency modes of the system (see equation (4.2)).

6. Conclusions. The numerical solution of overdetermined Sylvester equations (1.4) has applications in both the reduced-order modeling and the control of large dimensional systems as well as low-rank approximate solution of Lyapunov equations (1.1) and Sylvester equations (2.2). Our solution procedure involves the reduction of the original problem to a minimization of dimension at most $3k \times k$, followed by either a CG algorithm for the general case, or a PCG algorithm for minimizations (1.4) that are low-rank perturbations of a reduced-order general Sylvester equation (1.6). A CG algorithm requires $O(k^5)$ flops before convergence, while a PCG algorithm may require as few as $O(k^4)$ flops before convergence.

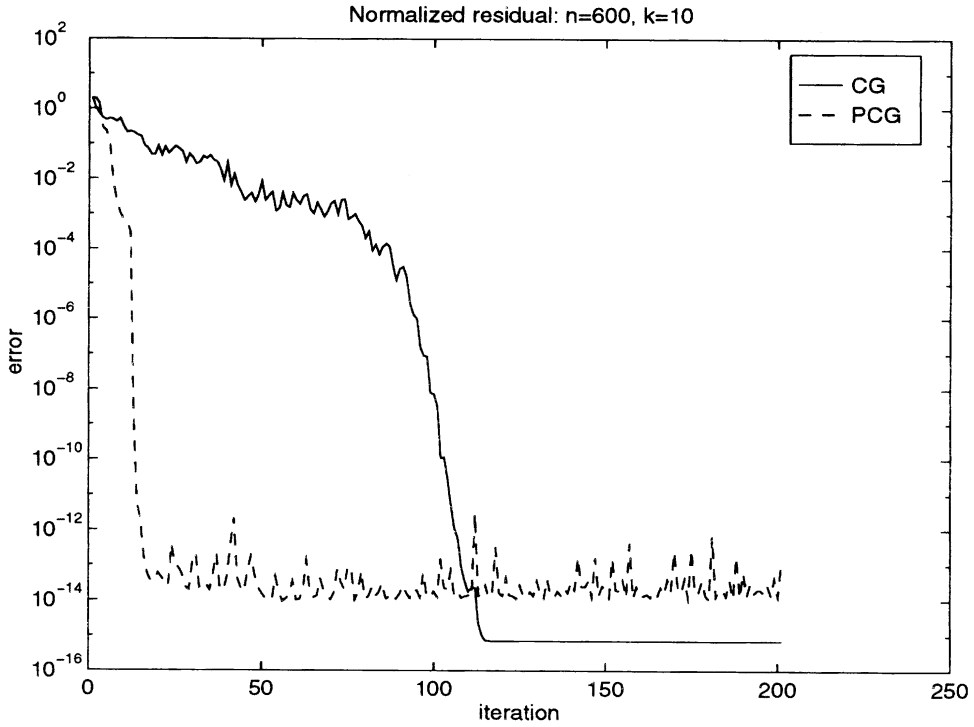


FIG. 5.2. *Residual plot:* $\delta = 0.1$, $\rho = 1$, $m = 1$.

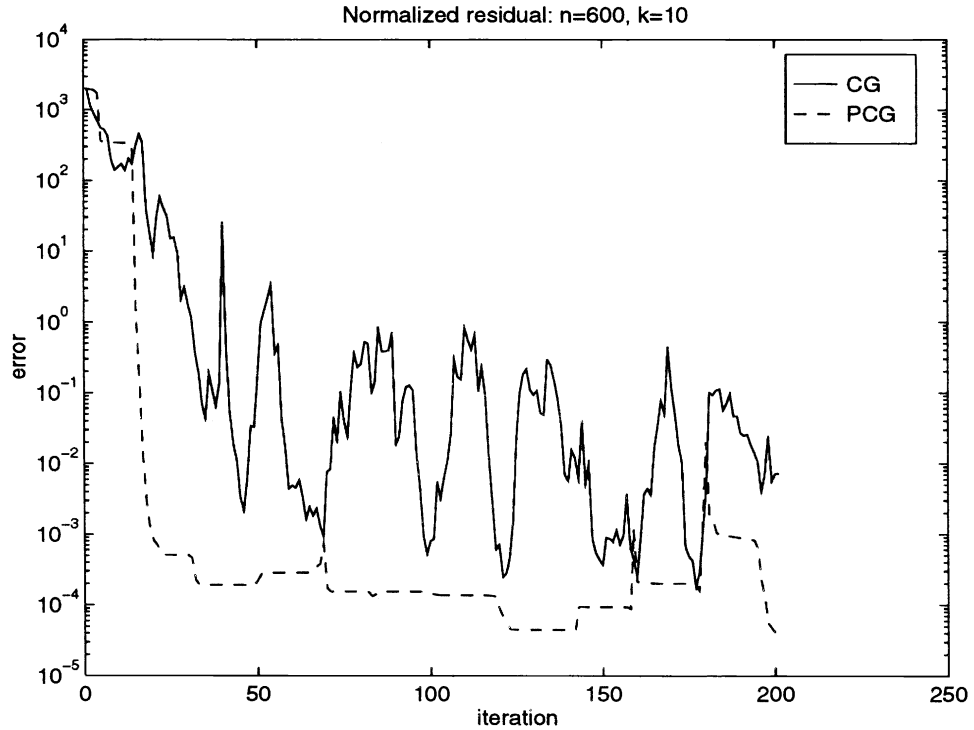


FIG. 5.3. *Residual plot:* $\delta = 10^{-3}$, $\rho = 10$, $m = 0.01$.

Acknowledgment. We wish to thank Gene Golub for suggesting the overdetermined Sylvester equation problem.

REFERENCES

[1] R. H. Bartels and G. W. Stewart, *Solution of the matrix equation $AX + XB = C$*, Comm. of the ACM, 15 (1972), pp. 820–826.

[2] D. S. Bernstein and D. C. Hyland, *The optimal projection equations for reduced-order state estimation*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 583–585.

[3] J. W. Brewer, *Kronecker products and matrix calculus in system theory*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 772–781.

[4] P. N. Brown, *A theoretical comparison of the Arnoldi and GMRES algorithms*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 58–78.

[5] R. Byers, *A LINPACK-style condition estimator for the equation $AX - XB = C$*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 926–928.

[6] P. Concus, G. H. Golub, and D. P. O'Leary, *A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations*, in Sparse Matrix Computations, J. R. Bunch and D. J. Rose, eds., Academic Press, New York, 1976, pp. 309–332.

[7] J. D. Gardiner, A. J. Laub, J. J. Amato, and C. B. Moler, *Solution of the Sylvester matrix equation $AXB^T + CXD^T = E$*, ACM Trans. Math. Software, 18 (1992), pp. 223–231.

[8] J. D. Gardiner, M. R. Wette, A. J. Laub, J. J. Amato, and C. B. Moler, *Algorithm* 705: *A Fortran-77 software package for solving the Sylvester matrix equation $AXB^T + CXD^T = E$*, ACM Trans. Math. Software, 18 (1992), pp. 232–238.

[9] J. A. George, M. T. Heath, and R. J. Plemmons, *Solution of large-scale sparse least squares problems using auxiliary storage*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 416–429.

[10] G. H. Golub, S. Nash, and C. Van Loan, *A Hessenberg-Schur method for the problem $AX + XB = C$*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 909–913.

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

[12] S. J. Hammarling, *Numerical solution of the stable, non-negative definite Lyapunov equation*, IMA J. Numer. Anal., 2 (1982), pp. 303–323.

[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. National Bureau of Standards, 49 (1952), pp. 409–436.

[15] A. S. Hodel, *Numerical Methods for the Solution of Large and Very Large, Sparse Lyapunov Equations*, Ph.D. thesis, Department of Electrical Engineering, University of Illinois at Urbana-Champaign, Champaign, IL, 1989.

[16] A. S. Hodel, *Least squares approximate solution of the Lyapunov equation*, in Proceedings of the 30th IEEE Conference on Decision and Control, Brighton, England, 1991, pp. 1619–1624.

[17] A. S. Hodel and K. Poolla, *Heuristic methods to the solution of very large, sparse Lyapunov and algebraic Riccati equations*, in Proceedings of the 27th IEEE Conference Decision and Control, Austin, TX, 1988, pp. 2217–2222.

[18] A. S. Hodel, R. B. Tenison, and K. Poolla, *Numerical solution of large Lyapunov equations by Approximate Power Iteration*, Linear Algebra Appl., 236 (1996), pp. 205–230.

[19] D. Y. Hu and L. Reichel, *Krylov subspace methods for the Sylvester equation*, Linear Algebra Appl., 172 (1992), pp. 283–313.

[20] D. C. Hyland and D. S. Bernstein, *The optimal projection equations for fixed-order dynamic compensation*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 1034–1037.

[21] M. Ilic, *New approaches to voltage monitoring and control*, IEEE Control Systems Magazine, 9 (1989), pp. 3–11.

[22] I. M. Jaimoukha and E. M. Kasenally, *Oblique projection methods for large scale model reduction*, SIAM J. Matrix. Anal. Appl., 16 (1995), pp. 602–627.

[23] I. M. Jaimoukha and E. M. Kasenally, *Krylov subspace methods for solving large Lyapunov equations*, SIAM J. Numer. Anal., 31 (1994), pp. 227–251.

[24] P. Lancaster, *Explicit solutions of linear matrix equations*, SIAM Review, 12 (1970), pp. 544–566.

[25] J. LaSalle and S. Lefschetz, *Stability of Liapunov's Direct Method*, Academic Press, New York, 1961.

[26] A. Lu, *Alternating Direction Implicit Iteration Solution of Lyapunov Equations*, Master's the-

sis, Department of Mathematics, University of Tennessee, Knoxville, TN, 1990.

[27] C. B. Moler and G. W. Stewart, *An algorithm for generalized matrix eigenvalue problems*, SIAM J. Numer. Anal., 10 (1973), pp. 241–256.

[28] B. C. Moore, *Principal component analysis in linear systems: Controllability, observability and model reduction*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 17–32.

[29] S. Richter and E. G. Collins Jr., *A homotopy algorithm for reduced order compensator design using the optimal projection equations*, in Proceedings of the 28th IEEE Conference on Decision and Control, Tampa, FL, 1989, pp. 506–511.

[30] S. Richter and A. S. Hodel, *Homotopy methods for the solution of general modified algebraic Riccati equations*, in Proceedings of the 29th IEEE Conference on Decision and Control, Honolulu, HI, 1990, pp. 971–976.

[31] Y. Saad, *Numerical solution of large Lyapunov equations*, in Signal Processing, Scattering and Operator Theory, and Numerical Methods, Progress in Systems and Control Theory Series, Vol. 5, M. A. Kaashoek, J. H. van Schuppen, and A. C. M. Ran, eds., Birkhauser, Boston, Cambridge, MA, pp. 401–410.

[32] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[33] R. J. Veillette, *Reliable Control of Decentralized Systems: an ARE-based H-infinity Approach*, Ph.D. thesis, Department of Electrical Engineering, University of Illinois at Urbana-Champaign, Champaign, IL, 1989.

[34] E. L. Wachspress, *Iterative solution of the Lyapunov matrix equation*, Appl. Math. Lett., 1 (1989), pp. 87–90.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, pp. 291–304, April 1997

003

EFFICIENT SOLUTION OF LINEARLY COUPLED LYAPUNOV EQUATIONS*

EMMANUEL G. COLLINS JR.† AND A. SCOTTEDWARD HODEL‡

Abstract. A numerical procedure is presented for the efficient solution of sets of linearly coupled matrix Lyapunov equations. Such equations arise in numerical continuation methods for the design of robust and/or low-order control systems.

Key words. Lyapunov equations, linear systems of equations

AMS subject classifications. 15A06, 15A24, 65F05

PII. S0895479894270178

1. Introduction.

The algebraic Riccati equation $A^TP+PA-PBP+C=0$ has played a central role in modern control theory. For example, the central computation in the synthesis of a linear-quadratical Gaussian (LQG) (globally $\mathcal{H}_2$ optimal) controller is the solution of two decoupled algebraic Riccati equations [1], [25]. It has also been shown more recently that the "central" controller yielding a closed-loop system with a specified $\mathcal{H}_\infty$ constraint can also be computed by first solving two decoupled algebraic Riccati equations [14], [16].

Unfortunately, both the LQG controller and central $\mathcal{H}_\infty$ controller have limited usefulness. First of all, both have dimension equal to the dimension of the design plant and hence may not be implementable due to constraints on the achievable throughput of the control processors. Also, it is well known that an LQG controller may be nonrobust and hence can destabilize or yield poor performance when implemented to control the physical plant [15], [31]. When properly designed, the full-order $\mathcal{H}_\infty$ controller can guarantee a certain measure of robust stability, but the design may be conservative because the $\mathcal{H}_\infty$ measure does not allow the incorporation of phase information regarding the uncertainty. In addition, $\mathcal{H}_\infty$ controllers do not adequately address the problem of robust performance.

As modern control has been extended to alleviate the deficiencies of the LQG and central $\mathcal{H}_\infty$ controllers, the computational requirements for control synthesis have become increasingly complex. For example, the design equations characterizing H_2 optimal reduced-order controllers appear as four nonlinearly coupled algebraic Riccati equations [17], [23]. Four coupled Riccati equations also characterize maximum entropy controllers [6], [7], [9], which have been seen to enable effectively the design of controllers that are robust with respect to frequency-dependent uncertainties [8], [9], [12], [13]. Mixed-norm $\mathcal{H}_2/\mathcal{H}_\infty$ controller synthesis, which can enable the design of multiple-objective controllers, similarly yields coupled Riccati equations [4], [19]. Recently developed control theory utilizes absolute stability theory (e.g., Popov stability theory) to synthesize controllers that are nonconservatively robust with respect

*Received by the editors June 24, 1994; accepted for publication (in revised form) by P. Van Dooren April 19, 1996. This work was supported in part by National Science Foundation grant ECS-9110083 and by Air Force Office of Scientific Research contract F49620-91-C-0019.
http://www.siam.org/journals/simax/18-2/27017.html
†Department of Mechanical Engineering, Florida A&M/Florida State University, Tallahassee, FL 32310 (ecollins@evax11.eng.fsu.edu).
‡Department of Electrical Engineering, 200 Broun Hall, Auburn University, Auburn, AL 36849 (scotte@eng.auburn.edu).

to real parametric uncertainty [20], [22]. This theory has also led to coupled Riccati equations. In fact, coupled Riccati equations appear in theories for optimal pole placement [21], optimal output feedback [26], optimal model reduction [10], [18], [24], [32], [33], and optimal reduced-order state estimation [5].

It is possible to solve sets of nonlinearly coupled Riccati equations using either continuation or homotopy algorithms [9], [10], [11], [30], [32], [33]. These algorithms are essentially prediction/correction schemes. The computation of the tangent directions for prediction or Newton steps for correction typically requires the numerical solution of sets of *linearly* coupled Lyapunov equations

$$A_i X_i + X_i {A_i}^T + \sum_{j=1}^{N} F_{ij}(X_j) + B_i = 0, \qquad i = 1, \ldots, N, \tag{1.1}$$

where A_i, X_i, and B_i are real $n_i \times n_i$ matrices, $B_i = {B_i}^T$, and F_{ij} is a symmetric linear operator; i.e., F_{ij} maps the set of $n_j \times n_j$ symmetric matrices into the set of $n_i \times n_i$ symmetric matrices.

Richter, Davis, and Collins [29] outline an algorithm for the solution of a single modified Lyapunov equation

$$AX + XA^T + B + F(X) = 0, \ X \in \mathbb{R}^{n \times n}, \tag{1.2}$$

where F is a symmetric linear operator. The basis of their algorithm is as follows: if the value of the perturbation function $F(X)$ were known a priori, then equation (1.2) reduces to the solution of a standard Lyapunov equation. Further, the dimension $\delta \triangleq \dim\left(\text{span}\left(F(\cdot)\right)\right)$ of the range space of the perturbation terms F is often quite small ($\delta \ll n^2$). With these features in mind, we may broadly summarize the algorithm of [29] in the following steps.

1. Select a set of basis functions $\phi_k \in \mathbb{R}^{n \times n}$ such that $F(X)$ can be written as a linear combination of ϕ_k for any value of X; i.e.,

$$F(X) = \sum_{k=1}^{\delta} \phi_k y_k \tag{1.3}$$

for appropriate scalars y_k, dependent on X.

2. Construct and solve a linear system of equations

$$(I + G)y = d$$

for the unknown parameters $y = \left[\begin{array}{ccc} y_1 & \ldots & y_\delta \end{array} \right]^T$ *without* first solving for the unknowns X.

3. Compute the value of the perturbation term $F(X)$ from equation (1.3) and substitute into equation (1.2).

4. Compute the unknown matrix X by the numerical solution of a standard Lyapunov equation.

The construction of the matrix G in step 2 requires the numerical solution of a set of Lyapunov equations

$$AX_i + X_i A^T + B_i = 0, \qquad i = 1, \ldots, \delta. \tag{1.4}$$

Since the equations (1.4) differ only in the right-hand side matrices $B_i, \ldots, B_\delta$, a single linear transformation $\tilde{A} \leftarrow T^{-1}AT$ is performed so that each of the δ Lyapunov

equations in (1.4) may be solved rapidly (in $O(n^2)$ flops versus $O(n^3)$ for, e.g., the Bartels–Stewart algorithm [2]). The purpose of this change of basis is strictly for the reduction of computational burden in the numerical solution of (1.4).

The above algorithm may be adapted in a straightforward fashion for the solution of sets of coupled Lyapunov equations (1.1). A preliminary procedure for this purpose is presented in [29]. A deficiency of this approach is that it fails to make use of the underlying structure in the coupling terms F_{ij}, $1 \le i, j \le N$, between the Lyapunov equations (1.1). For example, if the dependency structure of the set of equations (1.1) forms a directed, acyclic graph [28] (i.e., the block data dependency matrix, defined later in this paper, is upper triangular), then a maximally efficient algorithm should simply reduce to the solution of a series of N Lyapunov equations that are solved sequentially. However, the algorithm in [29] does *not* reduce in this manner. The present paper extends the algorithm of [29] to exploit the relative structure of the coupling terms in the coupled Lyapunov equations (1.1) and provides some examples of its application. More precisely, this algorithm exploits the block structure of the matrix G defined above while respecting the matrix-valued form of the original problem. The (usually) sparse block structure of the matrix G can be used to reduce the computational requirements (storage, flops) associated with the original algorithm of [29].

The remainder of this paper is organized as follows. We briefly discuss continuation methods in section 2 and further illustrate how linearly coupled Lyapunov equations arise in continuation algorithms for solving nonlinearly coupled Riccati equations. Section 3 then reviews the algorithm of [29] for the solution of linearly coupled Lyapunov equations, and section 4 extends the algorithm to make it efficiently exploit the relative structure of the coupling terms. Section 5 presents some examples to illustrate the results. Finally, section 6 gives the conclusions.

2. Coupled algebraic Riccati equations. Consider the set of N coupled algebraic Riccati equations

$$\mathcal{A}_i{}^T P_i + P_i \mathcal{A}_i - P_i B_i P_i + C_i + F_i(P_1, \ldots, P_N) = 0, \qquad i = 1, \ldots, N, \tag{2.1}$$

where

$$\mathcal{A}_i, B_i, C_i, P_i, \text{ and } \mathrm{Im}(F_i) \in \mathbb{R}^{n_i \times n_i}, \quad i = 1, \ldots, N,$$

F_i, $i = 1, \ldots, N$ are differentiable functions of $P_1, \ldots, P_N$, and $P_1, \ldots, P_N$ symmetric implies that

$$F_i(P_1, \ldots, P_N) = F_i(P_1, \ldots, P_N)^T, \ i = 1, \ldots, N.$$

Observe that a Lyapunov equation may be regarded as an algebraic Riccati equation with $B_i = 0$; hence, equation (2.1) encompasses the coupled algebraic Riccati and Lyapunov equations that arise in the motivational problems [3], [4], [5], [6], [7], [17], [18], [19], [20], [21], [22], [23], [24], [26].

For notational convenience, we shall rewrite the set of equations (2.1) as

$$\mathcal{A}^T P + P\mathcal{A} - PBP + C + F(P) = 0, \tag{2.2}$$

where $\mathcal{A} = \text{block-diag}(\mathcal{A}_1, \ldots, \mathcal{A}_N)$; the block diagonal matrices B, C, and P are defined similarly, and, with some abuse of notation,

$$\begin{aligned} F(P) &= \text{block-diag}(F_1(P_1, \ldots, P_N), \ldots, F_N(P_1, \ldots, P_N) \\ &= \text{block-diag}(F_1(P), \ldots, F_N(P)). \end{aligned}$$

In this form, the set of coupled modified algebraic Riccati equations (2.1) can be viewed as a *single* modified algebraic Riccati equation. Richter, Hodel, and Pruett [30] present a simple Newton-descent algorithm for the solution of general modified algebraic Riccati equations (2.2); however, this approach fails to exploit the block-diagonal structure of equation (2.2). In the simplest case in which the equations (2.1) are uncoupled, the individual solutions X_i may be obtained in $O(n_1^3 + \cdots + n_N^3)$ flops. However, since the algorithm of [30] ignores the underlying structure, this solution approach requires $O((n_1 + \cdots + n_N)^3)$ flops and so loses computational efficiency in this application.

Coupled algebraic Riccati equations of the form (2.1) can be solved through a continuation process or, if desired, a homotopy process; see [9], [10], [11], [30], [32], [33]. Continuation methods embed a problem $f(x) = 0$ into a parameterized family of problems

$$f(x, \lambda) = 0,$$

where the problems $f(x(\lambda), \lambda) = 0$, $\lambda \in [0, 1]$ are related by a continuous deformation parameterized in λ. For example, a candidate parameterization $P(\lambda)$ for equation (2.2) would be

$$\mathcal{A}^T P(\lambda) + P(\lambda)\mathcal{A} - P(\lambda)B(\lambda)P(\lambda) + C(\lambda) + F(P(\lambda)) + (\lambda - 1)E = 0, \tag{2.3}$$

where $E \triangleq F(P(0))$ and $P(0)$ satisfies the block-diagonal algebraic Riccati equation

$$\mathcal{A}^T P(0) + P(0)\mathcal{A} - P(0)B(0)P(0) + C(0) = 0.$$

Differentiation of equation (2.3) with respect to λ along the candidate continuation path may be performed in order to obtain a differential equation for $\dot{P}(\lambda) \triangleq dP(\lambda)/d\lambda$. The solution $P(1)$ of equation (2.2) (and, by consequence, the solution matrices $P_1, \ldots, P_N$ of equation (2.1)) may be obtained by the application of a numerical integration technique to compute $P(\lambda)$, $\lambda \in [0, 1]$. This approach has been used fruitfully in a number of studies on specific example design methodologies.

In this study, we shall consider the use of Newton and approximate-Newton descent strategies for the numerical solution of coupled algebraic Riccati equations (2.1). For this purpose, differentiate equation (2.3) with respect to λ to obtain

$$(\mathcal{A} - BP)^T \dot{P} + \dot{P}(\mathcal{A} - BP) - (P\dot{B}P + \dot{C} + E) + \frac{dF(P(\lambda))}{d\lambda} = 0, \tag{2.4}$$

where $\dot{P}(\lambda) \triangleq dP(\lambda)/d\lambda$. From matrix calculus properties, $dF/d\lambda$ will be linear in $\dot{P}(\lambda)$, and so equation (2.4) is a modified Lyapunov equation (1.2). The solution $\dot{P}(\lambda)$ of equation (2.4) may be used in a numerical integration scheme to follow the path $P(\lambda)$ from $P(0)$ (the solution of a standard algebraic Riccati equation) to $P(1)$ (a solution of the modified algebraic Riccati equation). Algorithms such as [9], [10], [11], [30] use a numerical integration technique in tandem with a Newton descent step in order to ensure tracking of the desired solution path $P(\lambda)$.

In many applications (e.g., [4], [5], [6], [7], [17], [18], [19], [20], [21], [22], [24], [26]), the set of coupled modified algebraic Riccati equations (2.1) involves perturbation functions $F_i(P)$ that (1) have a low-rank range space and (2) yield sparse data dependency graphs among the solutions $P_1, \ldots, P_N$. That is, the modified Lyapunov

equation (2.4) is more appropriately written as the set of linearly coupled Lyapunov equations

$$(\mathcal{A}_i - B_iP_i)^T\dot{P}_i + \dot{P}_i(\mathcal{A}_i - B_iP_i) - (P_i\dot{B}_iP_i + \dot{C}_i + E_i) + \sum_{j=1}^{N} F_{ij}(\dot{P}_j) = 0 \tag{2.5}$$

for appropriate linear operators F_{ij}, $i,j = 1,\ldots,N$, whose image forms a low-rank subspace of $\mathbb{R}^{n_i\times n_i}$. The numerical solution of sets of coupled Lyapunov equations is dealt with in section 3.

3. Parameterization of coupling. Consider the set of linearly coupled Lyapunov equations (1.1). The set of equations (1.1) may be written in a Kronecker product expansion through the application of the following theorem [29].

THEOREM 1. *Let F_{ij}, $1 \le i,j \le N$, be a set of linear functions that maps the set of N-tuples of symmetric matrices $(X_1,\ldots,X_N) \in (\mathbb{R}^{n_1\times n_1}\times\cdots\times\mathbb{R}^{n_N\times n_N})$ into itself. Then there exist integers $p_1,\ldots,p_N$ and corresponding matrices L_{ij}^k, $M_{ij}^k \in \mathbb{R}^{n_i\times n_k}$, $1 \le i,j \le N$, $1 \le k, \le p_i$ such that*

$$F_{ij}(X_j) = \sum_{k=1}^{p_i}\left(L_{ij}^k X_j {M_{ij}^k}^T + M_{ij}^k X_j {L_{ij}^k}^T\right). \qquad \square \tag{3.1}$$

The set of coupled Lyapunov equations (1.1) may be written as a single linear system of equations as follows. Let $\bar{b}_i = \mathrm{vec}(B_i)$ (the vector stack of the columns of the matrix B_i) and define

$$\bar{b} = \begin{bmatrix} \bar{b}_1^T & \cdots & \bar{b}_N^T \end{bmatrix}^T = \mathrm{vec}\left(\begin{bmatrix} B_1 & \cdots & B_N \end{bmatrix}\right).$$

Define $\bar{x}_1,\ldots,\bar{x}_N$, and $\bar{x}$ similarly in terms of $X_1,\ldots,X_N$. Let

$$\bar{A}_i = I_{n_i}\otimes A_i + A_i\otimes I_{n_i}, \quad i = 1,\ldots,N, \tag{3.2}$$

where $Y\otimes Z = \begin{bmatrix} y_{ij}Z \end{bmatrix}$ is the Kronecker product and define $\bar{n} = \sum_{i=1}^{N}\bar{n}_i$, $\bar{n}_i \triangleq n_i^2$. A consequence of Theorem 1 is that the set of Lyapunov equations (1.1) may be written as

$$(\bar{A} + \bar{F})\bar{x} = -\bar{b}, \tag{3.3}$$

where

$$\bar{A} = \text{block-diag}(\bar{A}_1,\ldots,\bar{A}_N) \in \mathbb{R}^{\bar{n}\times\bar{n}} \tag{3.4}$$

and $\bar{F} = \begin{bmatrix} \bar{F}_{ij} \end{bmatrix}$, where

$$\bar{F}_{ij} = \sum_{k=1}^{p_i} M_{ij}^k\otimes L_{ij}^k + L_{ij}^k\otimes M_{ij}^k \in \mathbb{R}^{\bar{n}_i\times\bar{n}_j}, \qquad 1 \le i,j \le N.$$

Observe that $\bar{F}_{ij}\mathrm{vec}(X) = \mathrm{vec}(F_{ij}(X))$ for all $X \in \mathbb{R}^{n_j\times n_j}$. While it is generally impractical to solve the set of equations (1.1) by the direct solution of (3.3), the latter equation may be used to determine dependency structure among the coupled solutions $X_1,\ldots,X_N$.

The matrix $\bar{F}$ frequently is of low rank; e.g., $\text{rank}(\bar{F}) = \bar{r} \ll \bar{n}$. In this event, it is possible to express

$$\bar{F} = \bar{C}\bar{D} \triangleq \begin{bmatrix} \bar{C}_1 \\ \vdots \\ \bar{C}_N \end{bmatrix} \begin{bmatrix} \bar{D}_1 & \cdots & \bar{D}_N \end{bmatrix}, \tag{3.5}$$

where $\bar{C}$, $\bar{D}^T \in \mathbb{R}^{\bar{n}\times\bar{r}}$ and such that

$$F_{ij}(X_j) = C_i(D_j(X_j)) \qquad \text{and} \qquad \text{vec}(F_{ij}(X_j)) = \bar{C}_i\bar{D}_j\bar{x}_j \tag{3.6}$$

for appropriately defined linear operators $C_1, \ldots, C_N$, $D_1, \ldots, D_N$, and for all $x_j \in \mathbb{R}^{\bar{n}_j}$.

Remark 1. For simplicity, we shall write the composition of two operators Ψ_1 and Ψ_2 as $\Psi_1\Psi_2(\cdot) \triangleq \Psi_1(\Psi_2(\cdot))$, or simply $\Psi_1\Psi_2$ when the context is clear. Similarly, we shall write the sum of two operators $(\Psi_1 + \Psi_2)(\cdot) = (\Psi_1 + \Psi_2) \triangleq \Psi_1(\cdot) + \Psi_2(\cdot)$. Thus, equation (3.6) may be written as $\text{vec}(F_{ij}(X_j)) = \text{vec}\,(C_iD_j(X_j))$.

The import of equation (3.6) is that should there exist a low-rank factorization (3.5), it is possible to represent the collective action of the N^2 linear operators F_{ij}, $1 \le i, j \le N$, in terms of $2N$ operators $C_1, \ldots, C_N, D_1, \ldots, D_N$, where the range of $(C_1, \ldots, C_N)$ composed with $(D_1, \ldots, D_N)$ is a rank $\bar{r}$ linear subspace. More precisely, let F be the linear operator defined as

$$\begin{aligned} F &: \mathbb{R}^{n_1\times n_1} \times \cdots \times \mathbb{R}^{n_N\times n_N} \to \mathbb{R}^{n_1\times n_1} \times \cdots \times \mathbb{R}^{n_N\times n_N} \\ &: (X_1, \ldots, X_N) \to \left(\sum_{j=1}^{N} F_{1j}(X_j), \ldots, \sum_{j=1}^{N} F_{Nj}(X_j) \right). \end{aligned} \tag{3.7}$$

Similarly, define

$$C : \mathbb{R}^{\bar{r}} \to \mathbb{R}^{n_1\times n_1} \times \cdots \times \mathbb{R}^{n_N\times n_N} : y \to (C_1(y), \ldots, C_N(y))$$

and

$$D : \mathbb{R}^{n_1\times n_1} \times \cdots \times \mathbb{R}^{n_N\times n_N} \to \mathbb{R}^{\bar{r}} : (X_1, \ldots, X_N) \to \sum_{j=1}^{N} D_j(X_j),$$

where

$$\text{vec}(C_j(y)) = \bar{C}_j y \text{ for } y \in \mathbb{R}^{\bar{r}} \tag{3.8}$$

and

$$D_j(X_j) = \bar{D}_j \text{vec}(X_j) \text{ for all } X_j \in \mathbb{R}^{n_j\times n_j}. \tag{3.9}$$

Then $F = CD$ has a range space of rank $\bar{r}$.

Richter, Hodel, and Collins [29] present a numerical procedure that makes use of this reduction in free parameters in order to solve the set of Lyapunov equations (1.1). The algorithm is summarized in the following definition and theorem.

DEFINITION 1. *Let ϕ be the linear operator*

$$\begin{aligned} \phi &: \mathbb{R}^{n_1\times n_1} \times \cdots \times \mathbb{R}^{n_N\times n_N} \to \mathbb{R}^{n_1\times n_1} \times \cdots \times \mathbb{R}^{n_N\times n_N} \\ &: (X_1, \ldots, X_N) \to (A_1X_1 + X_1{A_1}^T, \ldots, A_NX_N + X_N{A_N}^T). \end{aligned}$$

The action of the operator ϕ^{-1} on data $(B_1, \ldots, B_N)$ may be interpreted as the solution of the set of independent Lyapunov equations

$$A_i X_i + X_i A_i{}^T = B_i, \qquad 1 \le i \le N.$$

THEOREM 2 (see [29]). *Consider the set of coupled Lyapunov equations* (1.1). *Define the linear operators $C, D, \bar{C}$, and $\bar{D}$ as in equations* (3.6), (3.8), *and* (3.9). *Let*

$$\bar{d} = \mathrm{vec}\left(D\phi^{-1}(B_1, \ldots, B_N)\right) = \bar{D}\bar{A}^{-1}\mathrm{vec}\left(\begin{bmatrix} \bar{b}_1 & \cdots & \bar{b}_N \end{bmatrix}\right),$$

and define $\bar{G} = \bar{D}\bar{A}^{-1}\bar{C}$ as the matrix whose jth column may be computed as

$$\bar{g}_j = \mathrm{vec}\left(D\phi^{-1}C\left(e_j^{(\bar{r})}\right)\right),$$

where $e_j^{(\bar{r})}$ is the jth elementary vector of length $\bar{r}$. Then the set of equations (1.1) *has a solution $X_1, \ldots, X_N$ if and only if there exists a solution to*

$$\bar{y} = -(\bar{G}\bar{y} + \bar{d}). \tag{3.10}$$

In the event that a solution to (3.10) *exists, the set of solutions to* (1.1) *may be written as*

$$\mathcal{X} = \{(X_1, \ldots, X_N) : A_i X_i + X_i A_i{}^T + C_i(\bar{y}) + B_i = 0,\ \bar{y} = \bar{G}\bar{y} + \bar{d}\}. \tag{3.11}$$

Proof. Let $\bar{X}$ be the N-tuple of matrices $\bar{X} = (X_1, \ldots, X_n)$, and similarly define $\bar{B} = (B_1, \ldots, B_N)$. From (3.3), (3.6), (3.7), and Definition 1

$$(\phi + CD)(\bar{X}) = -\bar{B},$$

and so

$$\bar{X} = -\phi^{-1}\left(CD(\bar{X}) + \bar{B}\right).$$

Premultiply by D and define $\bar{y} = D(X_1, \ldots, X_N)$ to obtain

$$\bar{y} = -(D\phi^{-1}C(\bar{y}) + D\phi^{-1}(B_1, \ldots, B_N)) = -(D\phi^{-1}C(\bar{y}) + \bar{d}).$$

It is readily shown that $D\phi^{-1}C\bar{y} = \bar{D}\bar{A}^{-1}\bar{C}\bar{y}$ for all $\bar{y} \in \mathbb{R}^{\bar{r}}$, and the result follows. □

Remark 2. It follows from the reasoning in Theorem 2 that if ϕ is invertible, then $(I + \bar{G})$ is invertible if and only if there exists a unique solution $(X_1, \ldots, X_N)$ to the set of coupled Lyapunov equations (1.1). The applications listed in section 2 typically result in coupled Lyapunov equations with *stable* coefficient matrices A_i (i.e., $\lambda(A_i) \in$ left half-plane), and so the condition that ϕ be invertible is not overly restrictive.

The algorithm of [29] is summarized below.

PROCEDURE `LyapSetSolve`.

1. Compute the vector $\bar{y} = D(X_1, \ldots, X_N)$ *without* first computing the unknowns $X_1, \ldots, X_N$ as follows:
   (a) Transform the Lyapunov equations (1.1) into suitable bases for rapid solution (e.g., diagonal, block diagonal, tridiagonal).
   (b) Construct the vector $\bar{d} = D\phi^{-1}(B_1, \ldots, B_N)$.

(c) Construct the matrix $\bar{G} = \begin{bmatrix} \bar{g}_1 & \cdots & \bar{g}_N \end{bmatrix}$ as

$$\bar{g}_j = D\phi^{-1}\left(C_1 e_j^{(\bar{r})}, \ldots, C_N e_j^{(\bar{r})}\right),$$

where $C_1, \ldots, C_N$ are defined in (3.6) and $e_j^{(\bar{r})}$ is the jth elementary vector of length $\bar{r}$.

(d) Solve the linear system of equations

$$\bar{y} = -(\bar{G}\bar{y} + \bar{d}).$$

2. Solve the uncoupled set of Lyapunov equations

$$A_i X_i + X_i {A_i}^T + B_i + C_i(\bar{y}) = 0.$$

Observe that step 1c above requires the solution of Nr Lyapunov equations; the transformation in 1a is made in order to reduce the computational requirements in this step.

A weakness in the above approach is that it fails to detect exploitable structure in the data dependencies in the set of equations (1.1). For example, if the matrix $\bar{G}$ were (block) upper triangular (i.e., the data dependencies in the original problem (1.1) formed a directed, acyclic graph [28]), then the solutions $X_1, \ldots, X_N$ to (1.1) could be computed in reverse order using backward substitution; i.e., inversion of the entire matrix $(I + \bar{G})$ would be unnecessary. Similarly, if a subset of the equations (1.1) depended upon only a small number of other equations, then one could reduce $\bar{G}$ to a block upper triangular matrix through a straightforward manipulation of the corresponding rows of $\bar{G}$. It should be emphasized that much of the structure of the matrix $\bar{G}$ will be known based on the motivating design problem (see the example in section 5). We present our approach for identifying and exploiting the available structure of the matrix $\bar{G}$ in the next section.

4. Detection of dependency structure. In this section we discuss identification and the use of structure in the system of Lyapunov equations (1.1) so that the computational burden in the algorithm `LyapSetSolve` may be reduced. The principal area of attention is in algorithm steps 1b–1d. The issue that we address in this section is the failure of Procedure `LyapSetSolve` to exploit any underlying problem structure in the construction and solution of the linear system of equations

$$(I + \bar{G})\bar{y} = -\bar{d}$$

in steps 1c–1d.

From the definition of the reduced-dimension parameter vector $\bar{y} = D(X_1, \ldots, X_N)$ and the linearity of the operator D, we may assume without loss of generality that the vector $\bar{y}$ is partitioned and ordered such that $\bar{y} = \begin{bmatrix} \bar{y}_1^T & \cdots & \bar{y}_N^T \end{bmatrix}^T \in \mathbb{R}^{\bar{r}}$ with $\bar{y}_j \in \mathbb{R}^{r_j}$ a function of the unknown matrix X_j only. (It may be necessary for this purpose to increase the column dimension of $\bar{C}$, $\bar{D}^T$ to be larger than $\bar{r}$ (see (3.5)).) Conformably partition the matrix $\bar{G}$ and the vector $\bar{d}$; i.e.,

$$\begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_N \end{bmatrix} = -\left(\begin{bmatrix} \bar{G}_{11} & \cdots & \bar{G}_{1N} \\ \vdots & \ddots & \vdots \\ \bar{G}_{N1} & \cdots & \bar{G}_{NN} \end{bmatrix} \begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_N \end{bmatrix} + \begin{bmatrix} \bar{d}_1 \\ \vdots \\ \bar{d}_N \end{bmatrix} \right),$$

$$\bar{G}_{ij} \in \mathbb{R}^{r_i \times r_j}, \bar{d} \in \mathbb{R}^{r_j}.$$

We identify the structure of the matrix $\bar{G}$ via a block incidence matrix

$$\hat{G} = \begin{bmatrix} \hat{g}_{11} & \cdots & \hat{g}_{1N} \\ \vdots & \ddots & \vdots \\ \hat{g}_{N1} & \cdots & \hat{g}_{NN} \end{bmatrix},$$

where $\hat{G}_{ij} = 1$ if $\left\|\bar{G}\right\|_{ij} \neq 0$ and $\hat{g}_{ij} = 0$ otherwise. Suppose that some row i of $\hat{G}$ has only one nonzero entry $\hat{g}_{ij}$ off the diagonal. Then $\bar{y}_i$ is (directly) coupled only to $\bar{y}_j$ and may be written as

$$\bar{y}_i = (I + \bar{G}_{ii})^{-1} \left(\bar{d}_i - \bar{G}_{ij}\bar{y}_j\right).$$

Substitution of this identity in the remaining blocks of $\bar{G}$ and $\bar{d}$ decouples $\bar{y}_i$ from the remaining unknowns in $\bar{y}$ and thus reduces the associated computational burden. We state this approach more formally as follows.

PROCEDURE `SolveSys`($\hat{G}$,$\bar{G}(\cdot,\cdot)$, $\bar{d}(\cdot)$, k).

Inputs: $\hat{G}$: the block incidence matrix of $\bar{G}$ ($\hat{g}_{ij} \neq 0 \iff \bar{G}_{ij} \neq 0$).
$\bar{G}()$, $\bar{d}()$: $G(i,j)$ returns the block $\bar{G}_{ij}$, $\bar{d}(j)$ returns $\bar{d}_j$.
k: integer parameter used in problem reduction.

Outputs: Solution vector $\bar{y}$ satisfying $(I + \bar{G})\bar{y} = \bar{d}$.

1. Construct vectors I_r and I_c such that $I_c(j)$ = the total number of nonzero off-diagonal entries in column j of $\hat{G}$, and I_r is similarly defined in terms of the rows of $\hat{G}$. We shall refer to these vectors as the row/column total incidence vectors, respectively.
2. If possible, permute $\hat{G}$ to be block upper triangular. (A simplistic approach for this purpose would first permute $\hat{G} = P^T\hat{G}P$ so that $I_c(j)$ decreases monotonically and then permute "offending" rows/columns in order to maintain block upper triangular structure.)
3. For each block on the diagonal of $\hat{G}$:
   (a) Permute rows of $\hat{G}$ with fewer than k nonzero entries to the top of the block. (Use I_r to identify these rows.)
   (b) For each row permuted in step 3a (index = i)
      i. Identify indices $j_\nu, \nu = 1, \ldots, \nu_{max}$, where $\hat{g}_{ij_\nu} \neq 0$.
      ii. Identify indices $k_\sigma, \sigma = 1, \ldots, \sigma_{max}$, where $\hat{g}_{k_\sigma i} \neq 0$.
      iii. Set $\hat{g}_{k_\sigma i} = 0$, $\sigma = 1, \ldots, \sigma_{max}$ and update I_r, I_c.
      iv. For each pair (k_σ, j_ν), set $\hat{g}_{k_\sigma j_\nu} = 1$, $1 \leq \sigma \leq \sigma_{max}$, $1 \leq \nu \leq \nu_{max}$, update I_r, I_c, and record (k_σ, i, j_ν) in the index list. (In steps 3c and 3d below, the triple (k_σ, i, j_ν) indicates that $\bar{G}_{k_\sigma i}(I - \bar{G}_{ii})^{-1}\bar{G}_{ij_\nu}$ should be added to the $(k_\sigma j_\nu)$ block of $\bar{G}$ and that $(I - \bar{G}_{ii})^{-1}\bar{d}_i$ should be added to the vector block $\bar{d}_{k_\sigma}$.)
   (c) Permute remaining rows of the block to be block upper triangular, if possible.
   (d) For each block on the diagonal solve for the associated segment of $\bar{y}$ and back substitute.
4. Return the vector $\bar{y}$.

Remark 3. Step 2 differs from the permutation behavior of EISPACK routine `balanc` [27] in that it does not require that the leading/trailing blocks are upper triangular; the action of this step is to identify acyclic (one-way) block-data dependencies.

5. Numerical examples. The utility of the algorithm `SolveSys` is illustrated in the following examples.

Example 1. Maximum entropy design [6], [7] is a robust control system design technique that requires the numerical solution of the four coupled modified algebraic Riccati equations for matrix unknowns $P, Q, \hat{P}, \hat{Q}$ that satisfy

$$\begin{aligned} 0 &= {A_s}^T P + P A_s + R_1 - \left(B^T P + R_{12}\right)^T {R_2}^{-1} \left(B^T P + R_{12}\right) \\ &\quad + \sum_{i=1}^{n_\alpha} \alpha_i^2 {A_i}^T (P + \hat{P}) A_i, \end{aligned} \tag{5.1}$$

$$\begin{aligned} 0 &= A_s Q + Q {A_s}^T + V_1 - \left(QC^T + V_{12}\right) {V_2}^{-1} \left(QC^T + V_{12}\right)^T \\ &\quad + \sum_{i=1}^{n_\alpha} \alpha_i^2 A_i (Q + \hat{Q}) {A_i}^T, \end{aligned} \tag{5.2}$$

$$\begin{aligned} 0 &= (A_s - \left(QC^T + V_{12}\right) {V_2}^{-1} C) \hat{P} + \hat{P} (A_s - \left(QC^T + V_{12}\right) {V_2}^{-1} C)^T \\ &\quad + \left(B^T P + R_{12}\right)^T {R_2}^{-1} \left(B^T P + R_{12}\right), \end{aligned} \tag{5.3}$$

$$\begin{aligned} 0 &= (A_s - \left(B {R_2}^{-1} \left(B^T P + R_{12}\right)\right)^T \hat{Q} + \hat{Q} \left(A_s - \left(B {R_2}^{-1} \left(B^T P + R_{12}\right)\right)\right) \\ &\quad + \left(QC^T + V_{12}\right) {V_2}^{-1} \left(QC^T + V_{12}\right)^T \end{aligned} \tag{5.4}$$

(notation consistent with [9]). The matrices $A_1, \ldots, A_{n_\alpha}$ are typically rank 2 matrices that specify frequency-domain uncertainty in a nominal plant model. One recently proposed iterative algorithm for the numerical solution of the maximum entropy equations [9] requires the solution of four linearly coupled Lyapunov equations of the form

$$0 = {A_P}^T \mathcal{P} + \mathcal{P} A_P + R + \sum_{i=1}^{n_\alpha} \alpha_i^2 {A_i}^T (\mathcal{P} + \hat{\mathcal{P}}) A_i, \tag{5.5}$$

$$0 = A_Q \mathcal{Q} + \mathcal{Q} {A_Q}^T + V + \sum_{i=1}^{n_\alpha} \alpha_i^2 A_i (\mathcal{Q} + \hat{\mathcal{Q}}) {A_i}^T, \tag{5.6}$$

$$0 = {A_Q}^T \hat{\mathcal{P}} + \hat{\mathcal{P}} A_Q + \hat{R} + G_C \mathcal{Q} \hat{F} + \hat{F} \mathcal{Q} G_C + {H_P}^H \mathcal{P} + \mathcal{P} H_P, \tag{5.7}$$

$$0 = A_P \hat{\mathcal{Q}} + \hat{\mathcal{Q}} {A_P}^T + \hat{V} + G_B \mathcal{P} \hat{E} + \hat{E} \mathcal{Q} G_B + H_Q \mathcal{Q} + \mathcal{Q} {H_Q}^H \tag{5.8}$$

for unknowns $\mathcal{P}, Q, \hat{\mathcal{P}}, \hat{Q}$, where A_P, A_Q, $A_1, \ldots, A_{n_\alpha}$, $\hat{E}$, $\hat{F}$, G_B, G_C, H_P, H_Q, R, V, $\hat{R}$, $\hat{V} \in \mathbb{R}^{n \times n}$ and A_1, …, A_{n_α}, $\hat{E}$, $\hat{F}$, G_B, G_C, H_P, H_Q are low rank. Equations (5.5)–(5.8) are of the form (1.1) with $N = 4$. The coupling terms yield a block-incidence matrix

$$\hat{G} = \begin{bmatrix} \hat{g}_{11} & 0 & \hat{g}_{13} & 0 \\ 0 & \hat{g}_{22} & 0 & \hat{g}_{24} \\ \hat{g}_{31} & \hat{g}_{32} & 0 & 0 \\ \hat{g}_{41} & \hat{g}_{42} & 0 & 0 \end{bmatrix}.$$

Notice that the perturbations in (5.5)–(5.8) are functions of more than one unknown matrix; however, as observed in the previous section, the data vector $\bar{y}$ is easily constructed such that each segment $\bar{y}_i$ depends on exactly one unknown matrix.

Step 2 in Procedure `SolveSys` will be unable to permute $\hat{G}$ to block upper triangular form. However, observe that $\hat{G}$ has only two off-diagonal entries on rows 1 and

2. With $k \geq 1$, `SolveSys` step 3a identifies these two rows so that step 3b will zero out elements $\hat{g}_{31}, \hat{g}_{41}, \hat{g}_{32}$, and $\hat{g}_{42}$ after storing the triples $(3,1,3)$, $(4,1,3)$, $(3,2,4)$, and $(4,2,4)$ in the index list. The four triples correspond to substitutions in blocks $\bar{G}_{33}$, $\bar{G}_{43}$, $\bar{G}_{34}$, and $\bar{G}_{44}$, respectively, and so the corresponding elements of $\hat{G}$ are set nonzero with the result that

$$\hat{G} = \begin{bmatrix} \hat{g}_{11} & 0 & \hat{g}_{13} & 0 \\ 0 & \hat{g}_{22} & 0 & \hat{g}_{24} \\ 0 & 0 & * & * \\ 0 & 0 & * & * \end{bmatrix}$$

with *'d entries indicating modified nonzero entries. The matrix $\hat{G}$ is not further reducible by permutation, and so the vector segments $\bar{y}_3$ and $\bar{y}_4$ are computed as

$$\begin{bmatrix} \bar{y}_3 \\ \bar{y}_4 \end{bmatrix} = \left(\begin{bmatrix} \bar{G}_{33} - \bar{G}_{31}(I + \bar{G}_{11})^{-1}\bar{G}_{13} & -\bar{G}_{32}(I + \bar{G}_{22})^{-1}\bar{G}_{24} \\ \bar{G}_{41}(I + \bar{G}_{11})^{-1}\bar{G}_{13} & \bar{G}_{44} - \bar{G}_{42}(I + \bar{G}_{22})^{-1}\bar{G}_{24} \end{bmatrix} \right)^{-1} \times \begin{bmatrix} \bar{d}_3 + (I + \bar{G}_{11})^{-1}\bar{d}_1 \\ \bar{d}_4 + (I + \bar{G}_{22})^{-1} + \bar{d}_2 \end{bmatrix}.$$

These vector elements are then back substituted in order to obtain $\bar{y}_1$, $\bar{y}_2$, and the solution of the coupled set of Lyapunov equations (1.1) is obtained through Theorem 2.

Remark 4. The efficacy of our algorithm lies in the rapid solution of the matrix equation (3.10). The identification of block dependencies allows the careful implementation of both the construction of the matrix G and, in turn, the rapid computation of the parameter vector $\bar{y}$.

We further illustrate the algorithm behavior with the following numerical example.

Example 2. We tested our algorithm on a set of seven coupled Lyapunov equations (1.1), where $A_i = A \in \mathbb{R}^{7\times 7}$, $i = 1, \ldots, 7$ were tridiagonal with diagonal elements set to -2, superdiagonal elements set to 1, and subdiagonal elements set to -1, and $B_i = B = e_7 {e_7}^T$, $i-1, \ldots, 7$, where e_7 is the seventh elementary vector. For simplicity, the perturbations F_{ij} were selected as

$$F_{ij}(\cdot) = \begin{cases} e_i {e_j}^T(\cdot) e_j {e_i}^T, & i = j+1 \text{ and } (i,j) = (1,7), \\ 0 & \text{otherwise.} \end{cases}$$

(We discuss below the impact of higher-rank coupling terms on the computational cost of the algorithm.) The corresponding block-incidence matrix

$$\hat{G} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

is a periodic Markov probability transition matrix and hence irreducible by permutations only. Application of Procedure `SolveSys` with $k = 1$ to $\hat{G}$ yields an index list

of

$$\begin{bmatrix} 2 & 1 & 7 \\ 4 & 3 & 2 \\ 5 & 4 & 2 \\ 6 & 5 & 2 \\ 7 & 6 & 2 \end{bmatrix} \begin{array}{l} \text{delete entry at (2,1), add entry at (2,7)} \\ \text{delete entry at (4,3), add entry at (4,2)} \\ \text{etc.} \\ \vdots \end{array}$$

When the resulting block-incidence matrix is permuted so that the rows/columns appear in the order $\begin{bmatrix} 1 & 3 & 4 & 5 & 6 & 2 & 7 \end{bmatrix}$, then the matrix $\hat{G}$ is reduced to an upper triangular matrix

$$\hat{G}_{red} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

which is upper triangular except for the last two rows, which remain coupled. That is, where there were originally seven coupled blocks of the matrix G, now only two need to be solved simultaneously, and the segments of the parameter vector y that correspond to remaining blocks of $\hat{G}$ may be computed individually by back substitution.

For this example, the data vector is

$$d = \begin{bmatrix} 0.39338 \\ 5.2853 \cdot 10^{-6} \\ 2.4355 \cdot 10^{-5} \\ 1.5788 \cdot 10^{-4} \\ 1.0047 \cdot 10^{-3} \\ 6.5815 \cdot 10^{-3} \\ 0.045537 \end{bmatrix}$$

and the parameter vector is

$$\bar{y} = \begin{bmatrix} 0.41186 \\ 9.2950 \cdot 10^{-2} \\ 1.9400 \cdot 10^{-2} \\ 4.2056 \cdot 10^{-3} \\ 1.8823 \cdot 10^{-3} \\ 6.9743 \cdot 10^{-3} \\ 0.046984 \end{bmatrix}.$$

Remark 5. The computational savings of Procedure `SolveSys` are shown in Example 2 as follows. First, use of the block-incidence matrix $\hat{G}$ allows for the construction of the coupling matrix G through the solution of only $N = 7$ Lyapunov equations instead of $N^2 = 49$ Lyapunov equations; that is, the sparsity structure of $\hat{G}$ provides insight into the efficient construction of the matrix G. Further, once the matrix G is constructed, Procedure `SolveSys` obviates the need for the simultaneous inversion of the entire $N \times N$ matrix G, since only blocks 2 and 7 remain coupled in $\hat{G}_{red}$.

Remark 6. Since Example 2 used only rank 1 coupling terms for each Lyapunov equation, the savings in computational effort provided by `SolveSys` in this case is diminished from the general case. For example, suppose that the block-incidence matrix $\hat{G}$ was left unchanged but that the range space of the nontrivial coupling functions F_{ij} was a rank r subspace. Then the matrix $G \in \mathbb{R}^{Nr \times Nr}$ requires $O((Nr)^3)$ flops for inversion without the use of Procedure `SolveSys`. However, our procedure reduces this amount of work to $O(N(r^3))$ flops. That is, the greater the dimension of the range space of the coupling terms F_{ij}, the greater the savings provided by Procedure `SolveSys` over the original procedure presented in [29].

6. Conclusions. We have presented an algorithm for the solution of linearly coupled sets of Lyapunov equations (1.1). This algorithm is a refinement of an algorithm presented in [29], which failed to exploit data dependency structure in the coupling terms F_{ij}, $1 \leq i, j \leq N$. Our algorithm takes advantage of these data dependencies in order to reduce the computational burden in the solution of (1.1).

REFERENCES

[1] M. Athans, *The role and use of the stochastic linear-quadratic-gaussian problem in control system design*, IEEE Trans. Automat. Control, 16 (1971), pp. 529–552.

[2] R. H. Bartels and G. W. Stewart, *Solution of the matrix equation $AX + XB = C$*, Comm. of the ACM, 15 (1972), pp. 820–826.

[3] D. S. Bernstein, *Robust stability and performance via fixed-order dynamic compensation*, SIAM J. Control Optim., 27 (1989), pp. 389–406.

[4] D. S. Bernstein and W. M. Haddad, *LQG control with an H_∞ performance bound: A Riccati equation approach*, IEEE Trans. Automat. Control, 34 (1989), pp. 293–305.

[5] D. S. Bernstein and D. C. Hyland, *The optimal projection equations for reduced-order state estimation*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 583–585.

[6] D. S. Bernstein and D. C. Hyland, *The optimal projection/maximum entropy approach to designing low-order, robust controllers for flexible structures*, in Proceedings of the 24th IEEE Conference on Decision and Control, Tampa, FL, 1985, pp. 745–752.

[7] D. S. Bernstein and D. C. Hyland, *The optimal projection approach to robust, fixed-structure control design*, in Mechanics and Control of Large Flexible Structures, J. L. Junkins, ed., AIAA, Washington, DC, 1990, pp. 237–293.

[8] M. Cheung and S. Yurkovich, *On the robustness of MEOP design versus asymptotic LQG synthesis*, IEEE Trans. Automat. Control, 33 (1988), pp. 1061–1065.

[9] E. G. Collins Jr., L. D. Davis, and S. Richter, *Homotopy algorithm for maximum entropy design*, J. Guidance Control Dynamics, 17 (1994), pp. 311–321.

[10] E. G. Collins Jr., W. M. Haddad, S. Richter, and S. S. Ying, *An efficient, numerically robust homotopy algorithm for H_2 model reduction using the optimal projection equations*, Mathematical Modeling of Systems, Vol. 2, 1996, pp. 101–133.

[11] E. G. Collins Jr., W. M. Haddad, and S. Ying, *A homotopy algorithm for reduced-order dynamic compensation using the Hyland-Bernstein optimal projection equations*, J. Guidance Control Dynamics, 19 (1996), pp. 407–417.

[12] E. G. Collins Jr., J. A. King, and D. S. Bernstein, *Application of maximum entropy/optimal projection design synthesis to a benchmark problem*, J. Guidance Control Dynamics, 15 (1992), pp. 1094–1102.

[13] E. G. Collins Jr., D. J. Phillips, and D. C. Hyland, *Robust decentralized control laws for the ACES structure*, IEEE Control Systems Magazine, 11 (1991), pp. 62–70.

[14] J. Doyle, K. Glover, P. Khargonekar, and B. Francis, *State-space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 831–847.

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

[16] K. Glover and J. C. Doyle, *State-space formulae for all stabilizing controllers that satisfy an H^∞-norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.

[17] W. M. Haddad, *Robust Optimal Projection Control-System Synthesis*, Ph.D. thesis, Department of Mechanical Engineering, Florida Institute of Technology, Melbourne, FL, 1987.

[18] W. M. Haddad and D. S. Bernstein, *Combined L_2/H_∞ model reduction*, Internat. J. Control, 49 (1989), pp. 691–710.

[19] W. M. Haddad and D. S. Bernstein, *Generalized Riccati equations for the full and reduced-order mixed-norm H_2/H_∞ standard problem*, Systems Control Lett., 14 (1990), pp. 185–197.

[20] W. M. Haddad and D. S. Bernstein, *Parameter-dependent Lyapunov functions, constant real parameter uncertainty, and the Popov criterion in robust analysis and synthesis, part 1, part 2*, in Proceedings of the IEEE Conference on Decision and Control, Brighton, UK, 1991, pp. 2274–2279, 2617–2623.

[21] W. M. Haddad and D. S. Bernstein, *Controller design with regional pole constraints*, IEEE Trans. Automat. Control, AC-37 (1992), pp. 54–69.

[22] W. M. Haddad and D. S. Bernstein, *Parameter-dependent Lyapunov functions and the Popov criterion in robust analysis and synthesis*, IEEE Trans. Automat. Control, 40 (1995), pp. 536–543.

[23] D. C. Hyland and D. S. Bernstein, *The optimal projection equations for fixed-order dynamic compensation*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 1034–1037.

[24] D. C. Hyland and D. S. Bernstein, *The optimal projection equations for model reduction and the relationships among the methods of Wilson, Skelton, and Moore*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1201–1211.

[25] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.

[26] W. S. Levine and M. Athans, *On the determination of the optimal constant output feedback gains for linear multivariable systems*, IEEE Trans. Automat. Control, AC-15 (1970), pp. 44–48.

[27] B. N. Parlett and C. Reinsch, *Balancing a matrix for calculation of eigenvalues and eigenvectors*, Numer. Math., 13 (1969), pp. 293–304.

[28] E. M. Reingold, J. Nievergelt, and N. Deo, *Combinatorial Algorithms: Theory and Practice*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

[29] S. Richter, L. Davis, and E. G. Collins Jr., *Efficient computation of the solutions to modified Lyapunov equations*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 420–431.

[30] S. Richter, A. S. Hodel, and P. G. Pruett, *Homotopy methods for the solution of general modified algebraic Riccati equations*, IEE Proceedings Part D, 140 (1993), pp. 449–454.

[31] U. Shaked and E. Soroka, *On the stability robustness of the continuous-time LQG optimal control*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1039–1043.

[32] D. Zigic, L. T. Watson, E. G. Collins Jr., and D. S. Bernstein, *Homotopy approaches to the H_2 reduced order model problem*, J. Math. Systems Estimation Control, 3 (1993), pp. 173–205.

[33] D. Zigic, L. T. Watson, E. G. Collins Jr., and D. S. Bernstein, *Homotopy methods for solving the optimal projection equations for the H_2 reduced order model problem*, Internat. J. Control, 56 (1992), pp. 173–191.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, pp. 305–311, April 1997

004

SOME IMPROVEMENT OF OPPENHEIM'S INEQUALITY FOR M-MATRICES*

JIANZHOU LIU† AND LI ZHU†

Abstract. This paper improves Oppenheim's inequality as follows: if $A = (a_{ij})$ is an M-matrix and $B = (b_{ij})$ is an M-matrix or positive definite real symmetric matrix and A_k and B_k $(k = 1, 2, \ldots, n-1)$ are the $k \times k$ leading principal submatrices of A and B, respectively, then

$$\det(A \circ B) \geq a_{11} b_{11} \prod_{k=2}^{n} \left[b_{kk} \frac{\det A_k}{\det A_{k-1}} + \frac{\det B_k}{\det B_{k-1}} \left(\sum_{i=1}^{k-1} \frac{a_{ik} a_{ki}}{a_{ii}} \right) \right].$$

Key words. M-matrix, positive definite real symmetric matrix, determinant, Oppenheim's inequality

AMS subject classifications. 15A15, 15A48

PII. S0895479895296033

1. Introduction.

Let $\mathrm{R}^{m \times n}$ denote the set of $m \times n$ real matrices. Let S_n^+ denote the set of $n \times n$ positive definite real symmetric matrices. For $A = (a_{ij})$ and $B = (b_{ij}) \in R^{m \times n}$, the Hadamard product of A and B is the matrix $(a_{ij} b_{ij})$, which we denote by $A \circ B$. We write $A \geq B$ if $a_{ij} \geq b_{ij}$ for all i, j.

A real $n \times n$ matrix A is called an M-matrix if $A = sI - B$, where $s > 0, B \geq 0$, and $s > \rho(B)$, the spectral radius of B. Let M_n denote the set of $n \times n$ M-matrices. Let

$$Z^{n \times n} = \{A = (a_{ij}) \in R^{n \times n} : a_{ij} \leq 0 \text{ for all } i \neq j\}. \tag{1}$$

Suppose $A \in R^{n \times n}$, the comparison matrix $\mu(A) = (\mu_{ij})$, is defined by

$$\mu_{ij} = \begin{cases} -|a_{ij}|, & i \neq j, \\ |a_{ij}|, & i = j. \end{cases} \tag{2}$$

Let A be partitioned in the form

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where A_{11} is a nonsingular submatrix of A. $A/A_{11} = A_{22} - A_{21} A_{11}^{-1} A_{12}$ is called the Schur complement of A_{11} in A.

On the estimation of bounds for determinant, we have the following well-known classical result.

Oppenheim's inequality. If $A = (a_{ij})$ and $B = (b_{ij}) \in S_n^+$, then

$$\det(A \circ B) \geq \left(\prod_{i=1}^{n} a_{ii} \right) \det B. \tag{3}$$

* Received by the editors December 11, 1995; accepted for publication (in revised form) by T. Ando April 21, 1996.
http://www.siam.org/journals/simax/18-2/29603.html
† Department of Mathematics, Xiangtan University, Xiangtan, Hunan 411105, People's Republic of China.

Lynn [1] proved that inequality (3) holds for M-matrices and Fiedler and Ptak [2] gave a similar result when A is an M-matrix and B is a weakly diagonally dominant matrix.

In the paper, we shall obtain some inequalities which are stronger than inequality (3) for M-matrices and others.

LEMMA 1.1 (see [3]). *If A and $B \in M_n$, then $\mu(A \circ B) \in M_n$.*

LEMMA 1.2 (see [4]). *Let $A \in Z^{n\times n}$; then each of the following conditions is equivalent.*

(I) *A is an M-matrix.*

(II) *A has all positive diagonal elements, and there exists a positive diagonal matrix D such that AD is strictly diagonally dominant; that is,*

$$a_{ii}d_i > \sum_{j\neq i} |a_{ij}|d_j \qquad (i = 1, 2, \ldots, n).$$

(III) *There exists a positive diagonal matrix D such that $AD + DA^T \in S_n^+$.*

(IV) *All of the leading principal minors of A are positive.*

LEMMA 1.3 (see [4]). *If A and $B \in Z^{n\times n}, A \leq B$, and A is an M-matrix, then*

(I) *A^{-1} and B^{-1} exist and $A^{-1} \geq B^{-1} \geq 0$.*

(II) $\det B \geq \det A > 0$.

(III) *B is an M-matrix.*

LEMMA 1.4. *If A is a strictly diagonally dominant matrix with $a_{ii} > 0$ $(i = 1, 2, \ldots, n)$, then*

$$\det A \geq \det \mu(A) > 0. \tag{4}$$

Proof. By [1, Lemma 2.2], we have

$$|\det A| \geq \det \mu(A) > 0.$$

Furthermore, using the Geršgorin theorem, we obtain (4). □

LEMMA 1.5 (see [5]). *Let*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

If A_{11} is a nonsingular submatrix of A, then

$$\det(A/A_{11}) = \frac{\det A}{\det A_{11}}. \tag{5}$$

LEMMA 1.6 (see [6]). *If $A \in R^{n\times n}$, then*

$$\min \lambda \left(\frac{A + A^T}{2} \right) \leq \mathrm{Re}\lambda(A) \leq \max \lambda \left(\frac{A + A^T}{2} \right), \tag{6}$$

where $\lambda(A)$ denotes a characteristic root of matrix A.

2. Main results. In this section, for A and $B \in M_n$ we study Oppenheim's inequality.

THEOREM 2.1. *Let $A = (a_{ij})$ and $B = (b_{ij}) \in M_n$,A_k and $B_k (k = 1, 2, \ldots, n-1)$ be the $k \times k$ leading principal submatrices of A and B, respectively; then*

$$\det(A \circ B) \geq a_{11}b_{11} \prod_{k=2}^{n} \left[b_{kk} \frac{\det A_k}{\det A_{k-1}} + \frac{\det B_k}{\det B_{k-1}} \left(\sum_{i=1}^{k-1} \frac{a_{ik}a_{ki}}{a_{ii}} \right) \right]. \tag{7}$$

Proof. We prove (7) by induction on n. When $n = 1$, the inequality is understood as the trivial one, $\det(A \circ B) \geq a_{11}b_{11}$. We suppose that the result is valid for all $m \times m$ ($m < n$ and $n \geq 2$) M-matrices.

Let any $n \times n$ M-matrices A and B be partitioned:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & a_{nn} \end{bmatrix}, \qquad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & b_{nn} \end{bmatrix},$$

where A_{11} and $B_{11} \in M_{n-1}$.

For any $\varepsilon > 0$, let

$$\tilde{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{21}A_{11}^{-1}A_{12} + \varepsilon \end{bmatrix}, \qquad \tilde{B} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{21}B_{11}^{-1}B_{12} + \varepsilon \end{bmatrix};$$

then by the property of the Schur complement we shall get

$$\det \tilde{A} = \varepsilon \det A_{11} > 0, \ \det \tilde{B} = \varepsilon \det B_{11} > 0.$$

Therefore all of the leading principal minors of $\tilde{A}$ and $\tilde{B}$ are positive, so by Lemma 1.2 we have $\tilde{A}$ and $\tilde{B} \in M_n$.

By Lemma 1.2(II) then there exist two positive diagonal matrices D_1 and D_2 such that $\tilde{A}D_1$ and $\tilde{B}D_2$ are strictly diagonally dominant matrices, respectively. Thus

$$(\tilde{A} \circ \tilde{B})D_1D_2 = (\tilde{A}D_1) \circ (\tilde{B}D_2)$$

is a strictly diagonally dominant matrix. By Lemma 1.4, we have

$$\det(\tilde{A} \circ \tilde{B}) = \frac{1}{\det(D_1D_2)} \det[(\tilde{A}D_1) \circ (\tilde{B}D_2)] > 0. \tag{8}$$

Furthermore, using the property of the Schur complement and (4), (5), we have

$$\begin{aligned}
\det(\tilde{A} \circ \tilde{B}) &= \det(A_{11} \circ B_{11})[(A_{21}A_{11}^{-1}A_{12} + \varepsilon)(B_{21}B_{11}^{-1}B_{12} + \varepsilon) \\
&\quad - (A_{21} \circ B_{21})(A_{11} \circ B_{11})^{-1}(A_{12} \circ B_{12})] \\
&= \det(A_{11} \circ B_{11})[(a_{nn} - A/A_{11} + \varepsilon)(b_{nn} - B/B_{11} + \varepsilon) \\
&\quad - (A_{21} \circ B_{21})(A_{11} \circ B_{11})^{-1}(A_{12} \circ B_{12})] \\
&= \det(A_{11} \circ B_{11})[a_{nn}b_{nn} - (A_{21} \circ B_{21})(A_{11} \circ B_{11})^{-1}(A_{12} \circ B_{12})] \\
&\quad - \det(A_{11} \circ B_{11})[a_{nn}(B/B_{11}) + b_{nn}(A/A_{11}) - (A/A_{11})(B/B_{11}) \\
&\quad - (A_{21}A_{11}^{-1}A_{12})\varepsilon - (B_{21}B_{11}^{-1}B_{12})\varepsilon - \varepsilon^2] \\
&= \det(A \circ B) - \det(A_{11} \circ B_{11})[b_{nn}(A/A_{11}) + (A_{21}A_{11}^{-1}A_{12})(B/B_{11}) \\
&\quad - (A_{21}A_{11}^{-1}A_{12})\varepsilon - (B_{21}B_{11}^{-1}B_{12})\varepsilon - \varepsilon^2]. \qquad (9)
\end{aligned}$$

Letting $\varepsilon \to 0$ in (9), we have

$$\begin{aligned}
\det(A \circ B) &\geq \det(A_{11} \circ B_{11})[b_{nn}(A/A_{11}) + (A_{21}A_{11}^{-1}A_{12})(B/B_{11})] \\
&= \det(A_{11} \circ B_{11}) \left[b_{nn} \frac{\det A}{\det A_{n-1}} + \frac{\det B}{\det B_{n-1}} (A_{21}A_{11}^{-1}A_{12}) \right]. \qquad (10)
\end{aligned}$$

Since

$$\begin{bmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{n-1\ n-1} \end{bmatrix} \geq A_{11},$$

by Lemma 1.3, we have

$$A_{21}A_{11}^{-1}A_{12} \geq A_{21} \begin{bmatrix} a_{11}^{-1} & & 0 \\ & \ddots & \\ 0 & & a_{n-1\ n-1}^{-1} \end{bmatrix} A_{12} = \sum_{i=1}^{n-1} \frac{a_{in}a_{ni}}{a_{ii}}. \tag{11}$$

Combining (10), (11), and the inductive hypothesis, we obtain

$$\begin{aligned} \det(A \circ B) &\geq \det(A_{11} \circ B_{11}) \left[b_{nn} \frac{\det A}{\det A_{n-1}} + \frac{\det B}{\det B_{n-1}} \left(\sum_{i=1}^{n-1} \frac{a_{in}a_{ni}}{a_{ii}} \right) \right] \\ &\geq a_{11}b_{11} \prod_{k=2}^{n} \left[b_{kk} \frac{\det A_k}{\det A_{k-1}} + \frac{\det B_k}{\det B_{k-1}} \left(\sum_{i=1}^{k-1} \frac{a_{ik}a_{ki}}{a_{ii}} \right) \right]. \end{aligned}$$

□

COROLLARY 2.2. *If $A = (a_{ij})$ and $B = (b_{ij}) \in M_n$, then*

$$\det(A \circ B) \geq \left(\prod_{i=1}^{n} b_{ii} \right) \det A + \left(\prod_{i=1}^{n} a_{ii} \right) \det B \left(\prod_{k=2}^{n} \sum_{i=1}^{k-1} \frac{a_{ik}a_{ki}}{a_{ii}a_{kk}} \right). \tag{12}$$

Proof. Since all terms appearing in (12) are nonnegative, by Theorem 2.1 we have

$$\begin{aligned} \det(A \circ B) &\geq a_{11}b_{11} \prod_{k=2}^{n} \left[b_{kk} \frac{\det A_k}{\det A_{k-1}} + \frac{\det B_k}{\det B_{k-1}} \left(\sum_{i=1}^{k-1} \frac{a_{ik}a_{ki}}{a_{ii}} \right) \right] \\ &\geq a_{11}b_{11} \prod_{k=2}^{n} b_{kk} \frac{\det A_k}{\det A_{k-1}} + a_{11}b_{11} \prod_{k=2}^{n} \frac{\det B_k}{\det B_{k-1}} \left(\sum_{i=1}^{k-1} \frac{a_{ik}a_{ki}}{a_{ii}} \right) \\ &= \left(\prod_{i=1}^{n} b_{ii} \right) \det A + \left(\prod_{i=1}^{n} a_{ii} \right) \det B \left(\prod_{k=2}^{n} \sum_{i=1}^{k-1} \frac{a_{ik}a_{ki}}{a_{ii}a_{kk}} \right). \end{aligned}$$

□

Remark. Lynn [2] and Ando [7] have given the following result:

$$\det(A \circ B) + \det A \ \det B \geq \det A \prod_{i=1}^{n} b_{ii} + \det B \prod_{i=1}^{n} a_{ii} \tag{13}$$

for M-matrices A and B.

We point out that inequality (13) is better than (12).

In fact,

$$\begin{aligned} &\{\text{the left side} - \text{the right side of (12)}\} - \{\text{the left side} - \text{the right side of (13)}\} \\ &= \det B \prod_{i=1}^{n} a_{ii} \left\{ 1 - \frac{\det A}{\prod_{i=1}^{n} a_{ii}} - \prod_{k=2}^{n} \sum_{i=1}^{k-1} \frac{a_{ik}a_{ki}}{a_{ii}a_{kk}} \right\} \geq 0 \end{aligned}$$

or, equivalently, for any M-matrix $A = (a_{ij})$

$$1 - \frac{\det A}{\prod_{i=1}^{n} a_{ii}} - \prod_{k=2}^{n} \sum_{i=1}^{k-1} \frac{a_{ik}a_{ki}}{a_{ii}a_{kk}} \geq 0. \tag{14}$$

Here when $n = 1$ the expression $\prod_{k=2}^{n} \sum_{i=1}^{k-1} \frac{a_{ik}a_{ki}}{a_{ii}a_{kk}}$ is understood as 0.

To see (14), we may assume that $a_{11} = a_{22} = \cdots = a_{nn} = 1$. It is easy to see that (14) is true even with the equality sign for $n = 1, 2$.

Now assume that $n > 2$ and (14) is true for the case $n - 1$. Represent A, B as block forms

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & a_{nn} \end{bmatrix}, \qquad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & b_{nn} \end{bmatrix}.$$

Then by the Schur formula

$$\begin{aligned} &1 - \det A - \prod_{k=2}^{n} \sum_{i=1}^{k-1} a_{ik}a_{ki} \\ &= 1 - \det(A_{11})(1 - A_{21}A_{11}^{-1}A_{12}) - \prod_{k=2}^{n} \sum_{i=1}^{k-1} a_{ik}a_{ki} \\ &= \left\{ 1 - \det(A_{11}) - \prod_{k=2}^{n-1} \sum_{i=1}^{k-1} a_{ik}a_{ki} \right\} + \det(A_{11})\ A_{21}A_{11}^{-1}A_{12} \\ &\qquad + \left(1 - \sum_{i=1}^{n-1} a_{in}a_{ni} \right)\ \prod_{k=2}^{n-1} \sum_{i=1}^{k-1} a_{ik}a_{ki}. \end{aligned} \tag{15}$$

Since A_{11} is again an M-matrix, on the right side of (15) the first term is > 0 by the induction assumption and the second one is obviously > 0. Finally, as to the third term, by (11) we have

$$1 - \sum_{i=1}^{n-1} a_{in}a_{ni} \geq 1 - A_{21}A_{11}^{-1}A_{12} = \frac{\det A}{\det A_{11}} > 0.$$

This completes the induction.

THEOREM 2.3. *Let all the assumptions of Theorem* 2.1 *be satisfied. Then*

$$\det[\mu(A \circ B)] \geq a_{11}b_{11} \prod_{k=2}^{n} \left[b_{kk} \frac{\det A_k}{\det A_{k-1}} + \frac{\det B_k}{\det B_{k-1}} \left(\sum_{i=1}^{k-1} \frac{a_{ik}a_{ki}}{a_{ii}} \right) \right]. \tag{16}$$

Proof. Let $A, B, \tilde{A}$, and $\tilde{B}$ satisfy the assumptions of Theorem 2.1. By Lemma 1.1, we get $\mu(\tilde{A} \circ \tilde{B}) \in M_n$. Note that

$$\mu(A \circ B) = \begin{bmatrix} \mu(A_{11} \circ B_{11}) & -A_{12} \circ B_{12} \\ -A_{21} \circ B_{12} & a_{nn}b_{nn} \end{bmatrix},$$

$$\mu(\tilde{A} \circ \tilde{B}) = \begin{bmatrix} \mu(A_{11} \circ B_{11}) & -A_{12} \circ B_{12} \\ -A_{21} \circ B_{12} & (A_{21}A_{11}^{-1}A_{12} + \varepsilon)(B_{21}B_{11}^{-1}B_{12} + \varepsilon) \end{bmatrix}.$$

Then

$$0 < \det[\mu(\tilde{A} \circ \tilde{B})] = \det[\mu(A \circ B)] - \det[\mu(A_{11} \circ B_{11})][b_{nn}(A/A_{11})$$

$$+ (A_{21}A_{11}^{-1}A_{12})(B/B_{11}) - (A_{21}A_{11}^{-1}A_{12})\varepsilon - (B_{21}B_{11}^{-1}B_{12})\varepsilon - \varepsilon^2].$$

Now (16) can be proved in a similar manner as in the proof of Theorem 2.1. □

Remark. By Lemmas 1.2(II) and 1.4 we shall easily know that Theorem 2.3 is the improvement of Theorem 2.1.

3. A similar result. In this section, for $A \in M_n$ and $B \in S_n^+$, we study Oppenheim's inequality.

THEOREM 3.1. *If $A \in M_n$ and $B \in S_n^+$, then*

$$\det(A \circ B) > 0. \tag{17}$$

Proof. By Lemma 1.2(III), there exists a positive diagonal matrix D such that $AD + DA^T \in S_n^+$. By Lemma 1.6, we have

$$\begin{aligned} \mathrm{Re}\lambda[(AD) \circ B] &\geq \min \lambda \left[\frac{(AD) \circ B + (DA^T) \circ B}{2} \right] \\ &= \min \lambda \left[\frac{(AD) + (DA^T)}{2} \circ B \right] > 0. \end{aligned}$$

Therefore,

$$0 < \det[(AD) \circ B] = \det(A \circ B) \ \det D.$$

Thus we obtain (17). □

THEOREM 3.2. *If $A = (a_{ij}) \in M_n$ and $B = (b_{ij}) \in S_n^+$, then*

$$\det(A \circ B) \geq a_{11}b_{11} \prod_{k=2}^{n} \left[b_{kk} \frac{\det A_k}{\det A_{k-1}} + \frac{\det B_k}{\det B_{k-1}} \left(\sum_{i=1}^{k-1} \frac{a_{ik}a_{ki}}{a_{ii}} \right) \right]. \tag{18}$$

$\tilde{A}$ is an M-matrix and $\tilde{B}$ is positive definite by Theorem 3.1 $\det(\tilde{A} \circ \tilde{B}) > 0$. Now using a similar method of the proof of Theorem 2.1 we can prove Theorem 3.2.

COROLLARY 3.3. *If $A = (a_{ij}) \in M_n$ and $B = (b_{ij}) \in S_n^+$, then*

$$\det(A \circ B) \geq \left(\prod_{i=1}^{n} b_{ii} \right) \det A + \left(\prod_{i=1}^{n} a_{ii} \right) \det B \left(\prod_{k=2}^{n} \sum_{i=1}^{k-1} \frac{a_{ik}a_{ki}}{a_{ii}a_{kk}} \right). \tag{19}$$

Acknowledgments. The authors would like to thank Professor T. Ando and the referee for many suggestions which improved the paper. Particularly, Professor T. Ando has given a proof that inequality (13) is better than Corollary 2.2.

REFERENCES

[1] M. S. LYNN, *On the Schur product of the H-matrices and non-negative matrices and related inequalities*, Proc. Cambridge Philos., 60 (1964), pp. 425–431.

[2] M. Fiedler and V. Ptak, *Diagonally dominant matrices*, Czechoslovak Math. J., 92 (1967), pp. 420–433.
[3] K. Fan, *Inequalities for M-matrices*, Indag. Math., 26 (1964), pp. 602–610.
[4] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.
[5] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*, Academic Press, New York, 1979.
[6] M. Parodl, *La localisation des valeurs caracteristiques des matrices et ses applications*, Gauthier–Villars, Paris, 1959.
[7] T. Ando, *Inequalities for M-matrices*, Linear and Multilinear Algebra, 8 (1980), pp. 291–316.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, pp. 312–333, April 1997

005

HOMOTOPY METHOD FOR THE LARGE, SPARSE, REAL NONSYMMETRIC EIGENVALUE PROBLEM*

S. H. LUI†, H. B. KELLER‡, AND T. W. C. KWOK§

This paper is dedicated to Gene H. Golub on the occasion of his 65th birthday.

Abstract. A homotopy method to compute the eigenpairs, i.e., the eigenvectors and eigenvalues, of a given real matrix A_1 is presented. From the eigenpairs of some real matrix A_0, the eigenpairs of

$$A(t) \equiv (1-t)A_0 + tA_1$$

are followed at successive "times" from $t = 0$ to $t = 1$ using continuation. At $t = 1$, the eigenpairs of the desired matrix A_1 are found. The following phenomena are present when following the eigenpairs of a general nonsymmetric matrix:

- bifurcation,
- ill conditioning due to nonorthogonal eigenvectors,
- jumping of eigenpaths.

These can present considerable computational difficulties. Since each eigenpair can be followed independently, this algorithm is ideal for concurrent computers. The homotopy method has the potential to compete with other algorithms for computing a few eigenvalues of large, sparse matrices. It may be a useful tool for determining the stability of a solution of a PDE. Some numerical results will be presented.

Key words. eigenvalues, homotopy, parallel computing, sparse matrices, bifurcation

AMS subject classifications. 65F15, 65H17

PII. S0895479894273900

1. Introduction.

Given a real $n \times n$ matrix A, we wish to find some or all of its eigenvalues and eigenvectors. That is, we seek $\lambda \in \mathbb{C}$ such that

$$Ax = \lambda x$$

holds for nontrivial $x \in \mathbb{C}^n$. We call (x, λ) an eigenpair.

The QR algorithm (see Golub and Van Loan [9]) is generally regarded as the best sequential method for computing the eigenpairs. Briefly, the QR algorithm uses a sequence of similarity transformations to reduce a matrix to upper Hessenberg form. It then applies a sequence of Givens rotations from the left and right to reduce the size of the subdiagonal elements. When these elements are sufficiently small, the diagonal elements are taken to be approximations to the eigenvalues of the matrix. If the matrix is large and sparse, the QR algorithm suffers two serious drawbacks. In the reduction to Hessenberg form, the matrix usually loses its sparsity. Hence the algorithm requires the explicit storage of the entire matrix. This may pose a problem if the matrix is so large that not all of its entries can be accommodated within the main memory of the computer. A second drawback is that it is inherently a sequential algorithm due to the fact that Givens rotations must be applied sequentially. Bai and Demmel [3]

* Received by the editors September 7, 1994; accepted for publication (in revised form) by R. Freund April 29, 1996.
http://www.siam.org/journals/simax/18-2/27390.html
† Hong Kong University of Science & Technology, Department of Mathematics, Clear Water Bay, Kowloon, Hong Kong (shlui@uxmail.ust.hk). The work of this author was supported in part by RGC grant DAG92/93.SC16.
‡ California Institute of Technology, 217-50, Pasadena, CA 91125 (hbk@ama.caltech.edu).
§ Hong Kong University of Science & Technology, Department of Mathematics, Clear Water Bay, Kowloon, Hong Kong (mawkwok@uxmail.ust.hk).

somewhat circumvented the second problem by performing a "block" version of the QR algorithm. This improved version seems to work well on vector machines.

We now describe a homotopy method to compute the eigenpairs of a given matrix A_1. From the eigenpairs of some real matrix A_0, we follow the eigenpairs of

$$A(t) \equiv (1-t)A_0 + tA_1$$

at successive times from $t = 0$ to $t = 1$ using continuation. At $t = 1$, we have the eigenpairs of the desired matrix A_1. We call the evolution of an eigenpair as a function of time an eigenpath.

When A_1 is a real symmetric tridiagonal matrix with nonzero off-diagonal elements, a very successful homotopy method is known (see Li and Li [16] and Li, Zhang, and Sun [21]). The following phenomena, while absent in the symmetric tridiagonal case, are present for the general case:

- bifurcation,
- ill conditioning due to nonorthogonal eigenvectors.

The first can present computational difficulties if not handled properly. The homotopy method does not produce the Schur decomposition. Instead, it evaluates the eigenvalues and eigenvectors and hence is subject to the difficulty of ill conditioning.

Since the eigenpairs can be followed independently, this algorithm is ideal for parallel computers. We are primarily concerned with the case of a large, sparse, real matrix. We assume that all the nonzero entries of the matrix can be stored in each node of a parallel computer with distributive memory. Furthermore, we assume that the associated linear systems can be solved quickly, say, in $O(n^2)$ time.

As a simple illustration, we consider 2×2 matrices where the matrix A_0 is diagonal and whose elements are the diagonal elements of A_1:

$$A_0 = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}, \quad A_1 = \begin{bmatrix} a & d \\ c & b \end{bmatrix}.$$

The eigenvalues of $A(t)$ are

$$\frac{a+b \pm \sqrt{(a-b)^2 + 4t^2cd}}{2}.$$

Assuming $a \neq b$, three different situations arise (see Figure 1). In the first case, the two eigenvalues never meet for all t in $[0,1]$. In the second case, there is a double eigenvalue at some time $t \in (0,1]$ with the eigenpaths remaining real throughout. In the third case, there is a bifurcation point with the eigenpaths becoming a complex conjugate pair to the right of the bifurcation point. Typically, this is how complex eigenpaths arise from real ones. (Whenever a quantity is said to be complex, we mean it has a nontrivial imaginary component.) The situation for higher-dimensional matrices is similar except that an eigenpath can have more than one bifurcation point and the reverse of case three described above can occur (i.e., a complex conjugate pair of eigenpaths occur to the left of the bifurcation point and two real eigenpaths to the right). See Figure 2 for the eigenpaths of a random 10×10 matrix.

We now give a synopsis of the rest of the paper. In section 2, the homotopy method along with complex bifurcations will be presented. We will discuss some different types of bifurcations that may arise and identify the generic kind. We will derive an upper bound on the number of bifurcation points of all the eigenpaths. The numerical algorithm will be discussed in section 3. We will describe how to deal with

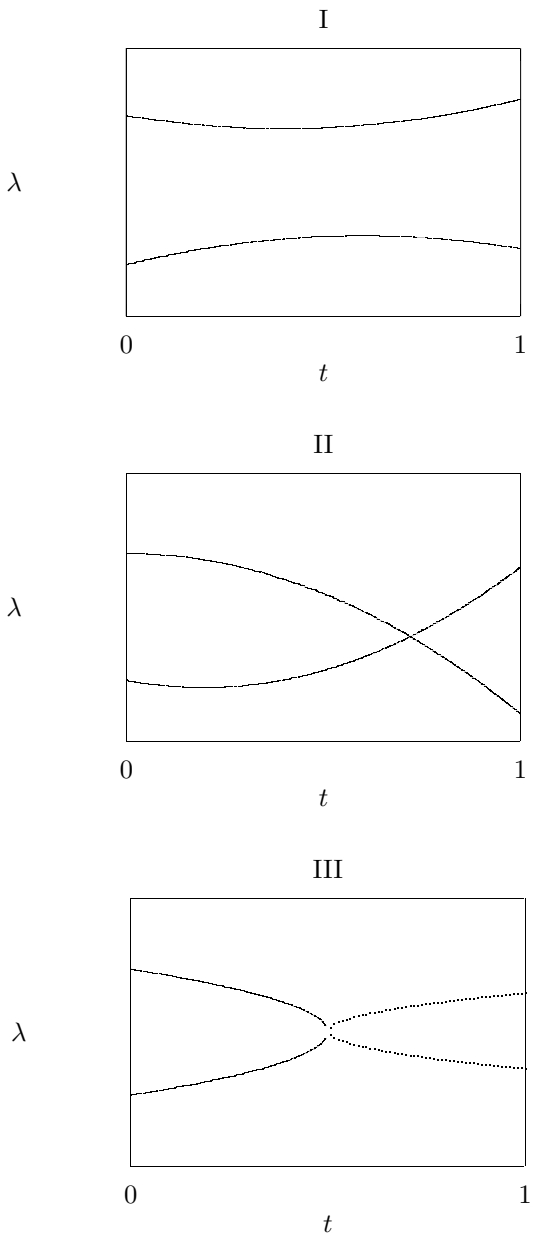


FIG. 1. *Eigenpaths of a* 2×2 *matrix. The dotted lines denote complex eigenpaths.*

bifurcations, how to choose the initial matrix, the selection of stepsizes etc. This will be followed by some numerical results. We will see that our homotopy method is impractical for dense matrices but has the potential to compete with other algorithms for finding a few eigenvalues of large, sparse matrices. Matrices of dimension 10^4 arising from the discretization of PDEs have been tested. In the final section, we recapitulate and suggest directions of further research.

Li, Zeng, and Cong [20] and Li and Zeng [19] have a very efficient homotopy

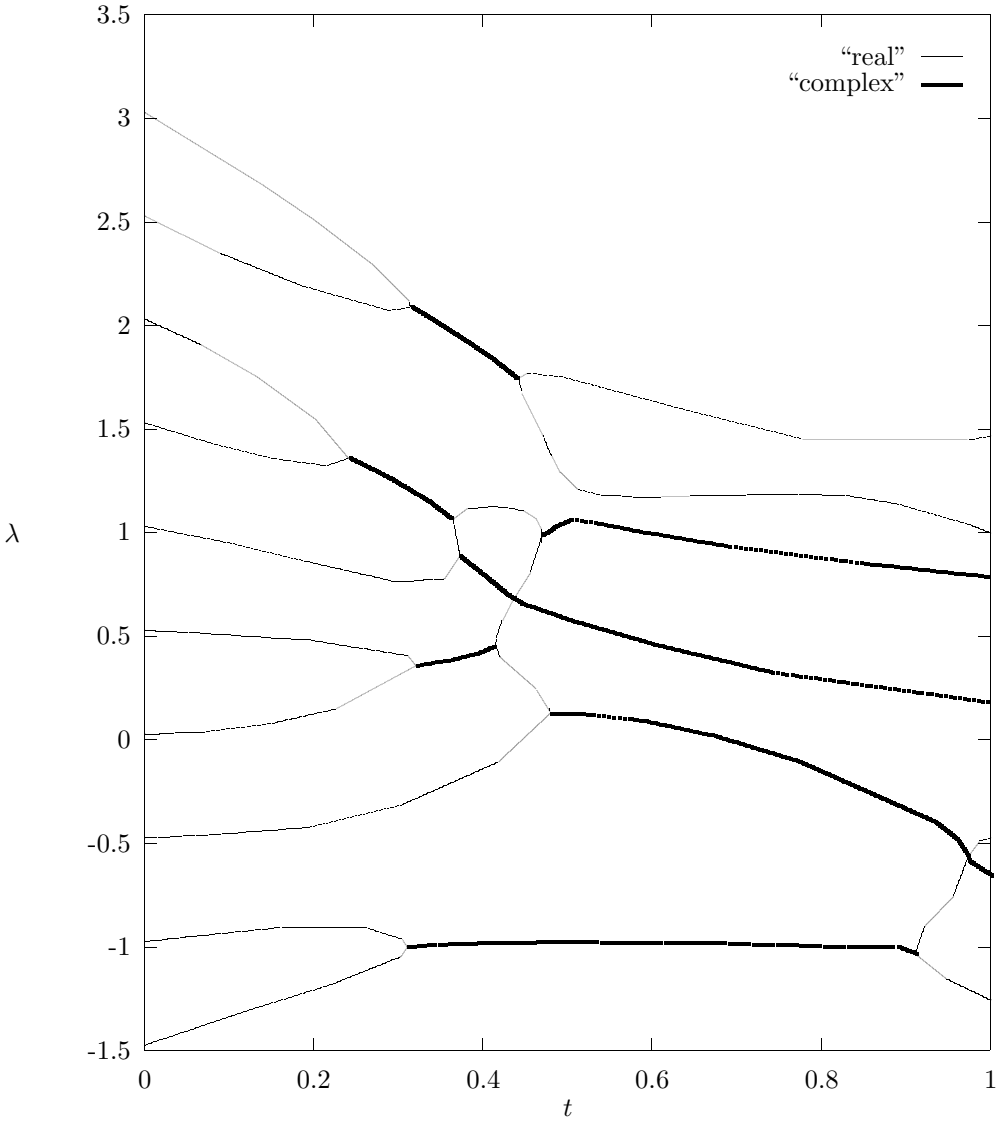


FIG. 2. *Eigenpaths of a random* 10×10 *matrix. Only one path of a complex conjugate pair of eigenpaths is shown.*

method for the dense matrix eigenvalue problem. For other approaches to the nonsymmetric eigenvalue problem see, for example, Cullum and Willoughby [5], Dongarra and Sidani [6], Saad [25], Shroff [28], Sorensen [29], Ruhe [24], and Bai, Day, and Ye [2]. The classic reference for the eigenvalue problem is the treatise by Wilkinson [30]. See also Saad [26] and Bai and Demmel [4] and the references therein.

Except for some of the numerical results, the work in this paper was completed by Lui [22]. In [20] Li, Zeng, and Cong proved Lemma A.1 (which they attribute to an unpublished work of Keller), which gives a necessary condition for a certain quantity $(\psi^*(G^0_{uu}\phi^2))$ to be nonzero. In this paper (Theorem 2), we give a necessary as well as

sufficient condition. Using analytic bifurcation theory, we identify the generic kinds of bifurcation which occur in following eigenpaths. We also give a bound on the number of bifurcation points in the eigenpaths. While the paper of Li, Zeng, and Cong addresses the dense eigenvalue problem, we address the complementary sparse case, although our algorithm has not had the same degree of success as theirs.

2. Homotopy method and complex bifurcation. In this section, we discuss some of the various phenomena that may arise on an eigenpath. Usually an eigenpath will be locally unique. That is, there are no other eigenpaths nearby. This can be characterized by a certain Jacobian being nonsingular. When this Jacobian is singular, bifurcation may occur. In other words, two or more eigenpaths may intersect at a point (u_0, t_0). Applying Henderson's work [10] on general analytic equations to our eigenvalue equations, we give a partial classification of some of the possible cases: simple quadratic fold, simple bifurcation point, simple cubic fold, and simple pitchfork bifurcation. We will show that the generic kind of bifurcation is the simple quadratic fold. In fact, the transition between real and complex eigenpaths (and vice versa) is via simple quadratic folds.

We first establish some notation. We use the superscripts ^{T} and * to denote the transpose and the complex conjugate transpose, respectively. The null and range spaces of a matrix are written as $\mathcal{N}()$ and $\mathcal{R}()$, respectively. The ith column of the identity matrix I is denoted by $\boldsymbol{e}_i$.

Given a real $n \times n$ matrix A_1, we form the homotopy

$$A(t) = (1-t)A_0 + tA_1, \quad 0 \le t \le 1, \tag{1}$$

where A_0 is a real matrix. We write the eigenvalue problem of $A(t)$ as

$$G(u,t) \equiv \begin{bmatrix} A(t)x - \lambda x \\ n(x) \end{bmatrix} = 0, \tag{2}$$

where u is the eigenpair (x, λ) of $A(t)$ and $n(x)$ is a normalization equation. In this paper, we take

$$n(x) = c^*x - 1,$$

where c is some fixed vector that is not orthogonal to x. The usual normalization $n(x) \equiv x^*x - 1$ is not differentiable, except at $x = 0$, and it only defines x up to a complex constant of magnitude one. We will always assume that every eigenvector x satisfies $c^*x \neq 0$; in section 3, we show how to choose c^*.

At this point, we make some remarks concerning the homotopy. It is known (Kato [12]) that the eigenvalues of $A(t)$ are analytic functions of t except at finitely many points where some eigenvalue may have an algebraic singularity. Away from these singularities, the eigenvectors *can be chosen* to be analytic functions of t. As we shall see, these singularities are typically encountered when an eigenvalue makes the transition from real to complex or vice versa.

Suppose an eigenpair u_0 is known at time t_0; i.e., $G(u_0, t_0) = 0$. We now describe how to obtain an eigenpair at a later time t_1. We must separate the discussion into different cases, depending on whether the Jacobian $G_u^0 \equiv G_u(u_0, t_0)$ is singular or not and on the nature of the singularity.

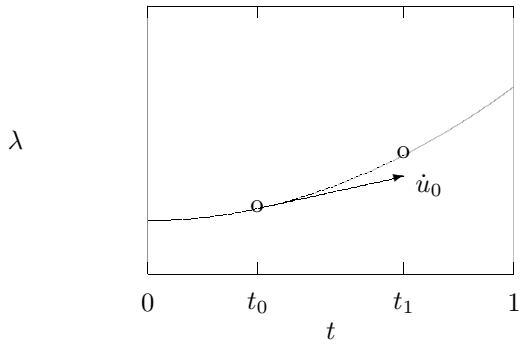


FIG. 3. *Euler–Newton continuation.*

2.1. Nonsingular Jacobian. When G_u^0 is nonsingular, the implicit function theorem tells us that locally about t_0 there is a unique solution $u(t)$ with $u(t_0) = u_0$. Differentiating (2) with respect to t and evaluating at t_0, we obtain

$$G_u^0 \dot{u}_0 + G_t^0 = 0,$$

where the dot denotes the t derivative and $G_t^0 \equiv G_t(u_0, t_0)$. Since G_u^0 is nonsingular, the above equation has a unique solution $\dot{u}_0$. To obtain the eigenpair at a later time t_1, we apply Newton's method to the equation $G(u, t_1) = 0$ with initial guess $u_0 + (t_1 - t_0)\dot{u}_0$. This is the Euler–Newton continuation method. The Euler step $(t_1 - t_0)\dot{u}_0$ is used to obtain the first Newton iterate (see Figure 3). Provided $t_1 - t_0$ is sufficiently small, the Newton iterates will converge quadratically to the eigenpair at t_1.

2.2. Singular Jacobian: Simple quadratic fold. Here we assume the eigenpair u_0 is real and

- G_u^0 has a one-dimensional null space spanned by, say, ϕ, and let ψ span the null space of $G_u^{0^T}$,
- $G_t^0 \notin \mathcal{R}(G_u^0)$,
- $a \equiv \psi^T(G_{uu}^0 \phi^2) \neq 0$.

Note that $G_{uu}^0 \phi^2$ is shorthand for $G_{uu}^0 \phi\phi$. The point (u_0, t_0) having the above properties is said to be a simple (real) quadratic fold point of equation (2). Pictorially, the real eigenpath is represented as the solid curve in Figure 4. Later, we will see that (1) λ_0 is an eigenvalue of $A(t_0)$ with algebraic multiplicity two and geometric multiplicity one and (2) $A'(t_0)x_0$ is not in the range of $[A(t_0) - \lambda_0 I, -x_0]$.

Since we can no longer use t to parametrize the solution, we employ the following pseudoarclength method due to Keller [13]. Augment (2) with the scalar equation

$$g(u, t, s) \equiv \phi^T \cdot (u - u_0) - (s - s_0) = 0.$$

This is the equation of a hyperplane whose unit normal is ϕ and is at a distance $s - s_0$ from u_0. Now define

$$F(u, t, s) \equiv \begin{bmatrix} G \\ g \end{bmatrix}. \tag{3}$$

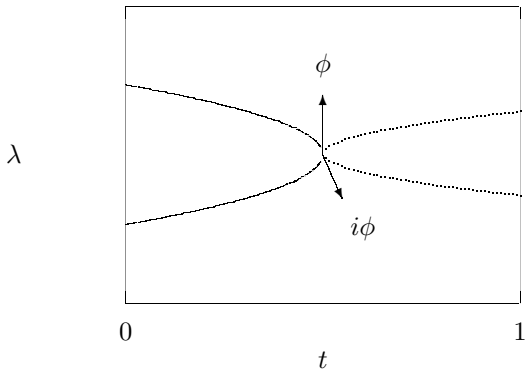


FIG. 4. *Complex conjugate pair of solutions on the opposite side of a simple real quadratic fold point. Dotted lines denote complex solutions.*

We immediately have $F(u_0, t_0, s_0) = 0$. It can be shown that the derivative of F with respect to (u, t) and evaluated at (u_0, t_0, s_0),

$$F^0_{(u,t)} = \begin{bmatrix} G^0_u & G^0_t \\ \phi^T & 0 \end{bmatrix}, \tag{4}$$

is nonsingular. Hence again by the implicit function theorem F has a locally unique solution $(u(s), t(s), s)$ with $u(s_0) = u_0$ and $t(s_0) = t_0$. In fact, the solution has the form

$$\begin{aligned} u(s) &= u_0 + \phi(s - s_0) + O(s - s_0)^2, \\ t(s) &= t_0 + \tau(s - s_0)^2 + O(s - s_0)^3, \end{aligned} \tag{5}$$

where

$$\tau = -\frac{1}{2}\frac{\psi^T(G^0_{uu}\phi^2)}{\psi^* G^0_t}.$$

From the definition of a simple quadratic fold, τ is well defined and nonzero. Note that $dt(s_0)/ds = 0$. We can apply the Euler–Newton continuation to the system $F = 0$ and follow the eigenpath around the fold point. Geometrically, the solution of $F = 0$ is the point at which the eigenpath punctures the hyperplane $g = 0$. Once around the fold point, t will begin to decrease. This is undesirable since our goal is to compute the eigenpair at $t = 1$. It turns out that a complex conjugate pair of eigenpaths will emerge to the right of the fold point. We now elaborate on this point.

Recall that a point $P_0 \equiv (u_0, t_0)$ is called a bifurcation point of the equation $G(u, t) = 0$ if in a neighborhood of P_0 there are at least two distinct branches of solutions $(u_1(s), t_1(s))$ and $(u_2(s), t_2(s))$ such that $u_i(s_0) = u_0$ and $t_i(s_0) = t_0$ for $i = 1, 2$. If at least one of these branches is complex, we will call P_0 a complex bifurcation point. When u_0 is real, (2) is a system of real equations. From the last paragraph, we know that locally about the point P_0 there is a unique path of *real* solutions. However, when considered as a system of equations over the complex numbers, Henderson and Keller [11] showed that P_0 is a complex bifurcation point with a complex conjugate pair of solutions on the opposite side of the real quadratic

fold (see Figure 4). Furthermore, the complex solutions have local expansions:

$$u(s) = u_0 + i\phi(s - s_0) + O(s - s_0)^2,$$
$$t(s) = t_0 - \tau(s - s_0)^2 + O(s - s_0)^3.$$

They are very similar in form to the real solution (5). Note that the tangent vector of the complex solution is a rotation of the tangent (ϕ) of the real solution. We can now use the Euler–Newton continuation with initial step in the direction $i\phi$ to find the complex eigenpairs at a later time.

The result of Henderson and Keller can be generalized to a complex quadratic fold point, i.e., $u_0 \in \mathbb{C}^{n+1}$, and satisfies the three properties outlined at the beginning of this section.

THEOREM 1 (Henderson [10]). *Let $G(u,t)$ be an analytic operator from $\mathbb{C}^{n+1} \times \mathbb{R}$ to $\mathbb{C}^{n+1}$. Let (u_0, t_0) be a simple quadratic fold point of $G(u,t) = 0$. Then in a small neighborhood of (u_0, t_0) there exist exactly two solution branches. They have the following expansions for small $|\epsilon|$:*

$$u_1(\epsilon) = u_0 + \epsilon e^{-i\alpha/2}\phi + O(\epsilon^2),$$
$$t_1(\epsilon) = t_0 - r\epsilon^2 + O(\epsilon^3),$$
$$u_2(\epsilon) = u_0 + i\epsilon e^{-i\alpha/2}\phi + O(\epsilon^2),$$
$$t_2(\epsilon) = t_0 + r\epsilon^2 + O(\epsilon^3),$$

where

$$re^{i\alpha} = \frac{\psi^*(G^0_{uu}\phi^2)}{2\psi^* G^0_t}.$$

2.3. Singular Jacobian: Simple quadratic bifurcation. Here, we assume the eigenpair u_0 is real and

- G^0_u has a one-dimensional null space spanned by, say, ϕ, and let ψ span the null space of $G^{0^T}_u$,
- $G^0_t \in \mathcal{R}(G^0_u)$,
- $a \neq 0$ and $b^2 - ac \neq 0$, where

$$a = \psi^T(G^0_{uu}\phi^2),$$
$$b = \psi^T(G^0_{uu}\phi\phi_0 + G^0_{ut}\phi),$$
$$c = \psi^T(G^0_{uu}\phi_0^2 + 2G^0_{ut}\phi_0),$$

and ϕ_0 is the unique solution of

$$G^0_u\phi_0 = -G^0_t \tag{6}$$

orthogonal to $\mathcal{N}(G^0_u)$.

The point (u_0, t_0) having the above properties is called a simple quadratic bifurcation point. In any small neighborhood of (u_0, t_0) there are exactly two distinct branches of solutions passing through the point (u_0, t_0) transcritically. If $b^2 - ac > 0$, then both branches are real. If $b^2 - ac < 0$, both branches are complex except at the point (u_0, t_0). See Henderson [10] for a more detailed discussion.

The tangent vectors of the two bifurcating branches can be computed and the Euler–Newton continuation can proceed as usual with these new directions. We will show that a simple quadratic bifurcation point is not likely to occur. Even if one existed, it would be transparent to a continuation method because it is highly unlikely that a numerical step would land exactly at the point.

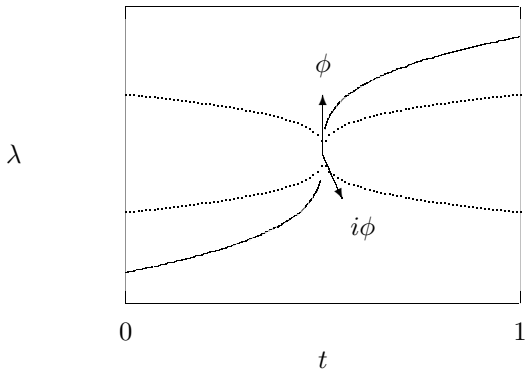


FIG. 5. *Cubic fold point.*

2.4. Singular Jacobian: Cubic fold point. Here, we assume the eigenpair u_0 is real and

- G_u^0 has a one-dimensional null space spanned by, say, ϕ, and let ψ span the null space of $G_u^{0^T}$,
- $G_t^0 \notin \mathcal{R}(G_u^0)$,
- $a \equiv \psi^T(G_{uu}^0\phi^2) = 0$,
- $\psi^T(G_{uu}^0\phi\phi_1) \neq 0$, where ϕ_1 is the unique solution of

$$G_u^0\phi_1 = -G_{uu}^0\phi^2 \tag{7}$$

  orthogonal to $\mathcal{N}(G_u^0)$.

The point (u_0, t_0) having the above properties is called a cubic fold point. It can be shown that (1) λ_0 is an eigenvalue of $A(t_0)$ with algebraic multiplicity three and geometric multiplicity one and (2) $A'(t_0)x_0$ is not in the range of $[A(t_0) - \lambda_0 I, -x_0]$. There is a unique branch of real solutions near (u_0, t_0) as well as a complex conjugate pair of solutions. See Figure 5. Cubic fold points are discussed, for example, in Yang and Keller [31] and Li and Wang [18]. Again, it will be seen that this case is not likely to occur in practice.

2.5. Singular Jacobian: Simple pitchfork bifurcation. Here, we assume the eigenpair u_0 is real and

- G_u^0 has a one-dimensional null space spanned by, say, ϕ, and let ψ span the null space of $G_u^{0^T}$,
- $G_t^0 \in \mathcal{R}(G_u^0)$,
- $a \equiv \psi^T(G_{uu}^0\phi^2) = 0$,
- $\psi^T(G_{uu}^0\phi\phi_1) \cdot \psi^T(G_{uu}^0\phi_0\phi + G_{ut}^0\phi) \neq 0$, where ϕ_0 and ϕ_1 were defined in (6) and (7).

The point (u_0, t_0) having the above properties is called a simple pitchfork bifurcation point. On one side of the point there are three real solutions. On the other side there is one real solution and a complex conjugate eigenpair. The situation is depicted in Figure 6. See Henderson [10] for a more detailed discussion.

2.6. Generic singular Jacobians. In the previous sections, we discussed four cases where the Jacobian G_u^0 has a one-dimensional null space. This list is of course not exhaustive. We will now see that of all the singularities only one, the simple quadratic fold, is likely to arise in the course of a calculation. The others are nongeneric.

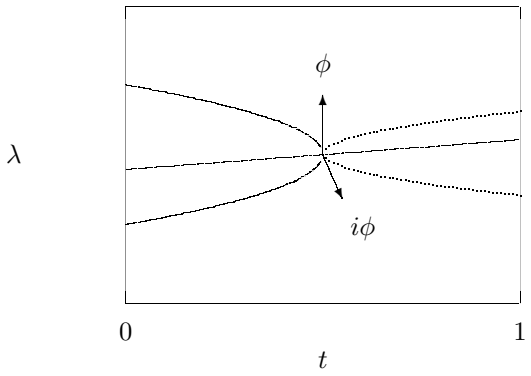


FIG. 6. *Simple pitchfork bifurcation.*

It is clear that of all the singular $n \times n$ matrices those with a one-dimensional null space are generic. Of the four cases considered, all but the first are nongeneric because they have nongeneric conditions $\psi^T(G^0_{uu}\phi^2) = 0$ and/or $G^0_t \in \mathcal{R}(G^0_u)$. The next result characterizes the generic singular Jacobian G^0_u.

THEOREM 2. *Let G be defined as in equation* (2). *Suppose for $(u_0, t_0) \in \mathbb{C}^{n+1} \times \mathbb{R}$, $G(u_0, t_0) = 0$ and G^0_u is singular with a one-dimensional null space. Let ϕ and ψ be spanning vectors for $\mathcal{N}(G^0_u)$ and $\mathcal{N}(G^{0^*}_u)$, respectively. Then $\psi^*(G^0_{uu}\phi^2) \neq 0$ iff λ_0 is an eigenvalue of $A^0 \equiv A(t_0)$ of algebraic multiplicity two and geometric multiplicity one.*

Proof. From (2), we obtain

$$G^0_u = \begin{bmatrix} A^0 - \lambda_0 I & -x_0 \\ c^* & 0 \end{bmatrix} \in \mathbb{C}^{n+1 \times n+1}.$$

Partition the null vectors as

$$\phi = \begin{bmatrix} h \\ \nu \end{bmatrix}, \quad \psi^* = [p^*, \mu],$$

where $h, p \in \mathbb{C}^n$ and $\nu, \mu \in \mathbb{C}$. By a direct calculation, we get

$$\psi^*(G^0_{uu}\phi^2) = -2\nu p^* h. \tag{8}$$

We rewrite the equation $\psi^* G^0_u = 0$, using the definitions of ψ^* and G^0_u, as

$$[p^*(A^0 - \lambda_0 I) + \mu c^*, \ -p^* x_0] = 0. \tag{9}$$

Taking the dot product of the first n components of the above vector with x_0, we obtain

$$p^*(A^0 - \lambda_0 I)x_0 + \mu c^* x_0 = 0.$$

Since $c^* x_0 = 1$,

$$\mu = 0. \tag{10}$$

The following two cases are the only possible ones in which $\dim \mathcal{N}(G^0_u) = 1$.

Case 1: λ_0 is an eigenvalue of A^0 with algebraic multiplicity $m \geq 2$ and geometric multiplicity one. Let

$$J \equiv Q^{-1}(A^0 - \lambda_0 I)Q = \left[\begin{array}{ccccc|c} 0 & 1 & & & & \\ & 0 & \ddots & & & \\ & & \ddots & 1 & & \\ & & & 0 & & \\ \hline & & & & & J_2 \end{array}\right] \tag{11}$$

be a Jordan form of $A^0 - \lambda_0 I$ where J_2 is nonsingular of dimension $n - m$ and x_0 is the first column of the matrix Q of principal (generalized) eigenvectors. Note that G_u^0 is similar to

$$\left[\begin{array}{cc} J & -\boldsymbol{e}_1 \\ c^*Q & 0 \end{array}\right].$$

Now from (9) and (10), we have

$$\begin{aligned} 0 &= p^*(A^0 - \lambda_0 I) \\ &= p^* Q J Q^{-1}. \end{aligned}$$

Let $y^* = p^*Q$. Then

$$y^* J = 0.$$

Thus from (11), we can take y^* to be $\boldsymbol{e}_m^*$.

From

$$G_u^0 \left[\begin{array}{c} h \\ \nu \end{array}\right] = 0,$$

we get

$$(A^0 - \lambda_0 I)h = \nu x_0. \tag{12}$$

Using (11) in the above, we obtain

$$QJQ^{-1}h = \nu x_0,$$

which implies that

$$Jw = \nu Q^{-1} x_0 = \nu \boldsymbol{e}_1,$$

where $w = Q^{-1}h$. From (11), we obtain the solutions $w = \alpha \boldsymbol{e}_1 + \nu \boldsymbol{e}_2$, where α is any complex number. Hence $y^* w = \nu \delta_{m2}$. Finally, from (8),

$$\begin{aligned} \psi^*(G_{uu}^0 \phi^2) &= -2\nu (p^*Q)(Q^{-1}h) \\ &= -2\nu y^* w \\ &= -2\nu^2 \delta_{m2}. \end{aligned}$$

Note that $\nu \neq 0$ since otherwise $w = \alpha \boldsymbol{e}_1$, which implies $h = \alpha x_0$. Since $c^*h = 0$ and $c^*x_0 = 1$, we must have $\alpha = 0$. We have reached a contradiction that ϕ is the zero vector. Hence $\psi^*(G_{uu}^0 \phi^2)$ is nonzero iff $m = 2$.

Case 2: λ_0 is an eigenvalue of A^0 with algebraic multiplicity $m \geq 2$ and geometric multiplicity two. Let

$$J \equiv Q^{-1}(A^0 - \lambda_0 I)Q = \left[\begin{array}{cc|c} J_1 & 0 & \\ 0 & J_2 & \\ \hline & & J_3 \end{array}\right] \tag{13}$$

be a Jordan form of $A^0 - \lambda_0 I$ where J_1 and J_2 are Jordan blocks of sizes m_1 and m_2, respectively, with $m_1 + m_2 = m$; J_3 is nonsingular and of dimension $n - m$; and x_0 is the first column of the matrix Q of principal eigenvectors. J_1 and J_2 have zeros on the diagonal. If J_1 is diagonal then, as before, we have from (12),

$$Jw = \nu \boldsymbol{e}_1,$$

where $w = Q^{-1}h$. From the form of J, it is clear that $\nu = 0$. Hence

$$\psi^*(G^0_{uu}\phi^2) = -2\nu p^* h = 0.$$

Finally, if J_1 is a nondiagonal Jordan block so that $m_1 > 1$, then J has at least two linearly independent left null vectors ($\boldsymbol{e}^*_{m_1}$ and $\boldsymbol{e}^*_m$). This implies that G^0_u has at least two linearly independent left null vectors ($[\boldsymbol{e}^*_{m_1}Q^{-1}, 0]$ and $[\boldsymbol{e}^*_m Q^{-1}, 0]$). (For example,

$$[\boldsymbol{e}^*_m Q^{-1}, 0]\, G^0_u = [\boldsymbol{e}^*_m Q^{-1}, 0] \left[\begin{array}{cc} Q & 0 \\ 0 & 1 \end{array}\right] \left[\begin{array}{cc} J & -\boldsymbol{e}_1 \\ c^*Q & 0 \end{array}\right] \left[\begin{array}{cc} Q^{-1} & 0 \\ 0 & 1 \end{array}\right] = 0$$

since $m > 1$ and $\boldsymbol{e}^*_m$ is a left null vector of J.) This contradicts the assumption that $\dim \mathcal{N}(G^0_u) = 1$.

Note that if λ_0 is an eigenvalue of A^0 of geometric multiplicity greater than two, it can be checked that the dimension of the null space of G^0_u is at least two. We have established the claim of the theorem. □

See also Li, Zeng, and Cong [20].

The fact that the generic case of a singular G^0_u occurs when λ_0 is an eigenvalue of A^0 of algebraic multiplicity two and geometric multiplicity one may seem surprising. We now attempt to give an intuitive explanation. Let X be the set of $n \times n$ matrices which have λ_0 as an eigenvalue of algebraic multiplicity two. Suppose A is a member of X. Now $A - \lambda_0 I$ can be similarly transformed to one of

$$\left[\begin{array}{cc|c} 0 & 1 & \\ 0 & 0 & \\ \hline & & J_1 \end{array}\right] \quad \text{or} \quad \left[\begin{array}{cc|c} 0 & 0 & \\ 0 & 0 & \\ \hline & & J_2 \end{array}\right],$$

where J_1 and J_2 are some nonsingular matrices. The rank of the left and right matrices are $n - 1$ and $n - 2$, respectively. Hence in the space X the matrix $A - \lambda_0 I$ with geometric multiplicity one (i.e., similar to the left matrix) is generic.

Using the notation of Theorem 2, we can show the following corollary.

COROLLARY 1. *Suppose $n \geq 4$ and $\mathcal{N}(G^0_u)$ is one dimensional. Then, generically, λ_0 is an eigenvalue of A^0 of algebraic multiplicity two and geometric multiplicity one and is real.*

Proof. Let X_r be the set of real $n \times n$ matrices with one real eigenvalue of algebraic multiplicity two and geometric multiplicity one and all other eigenvalues simple and let X_i be the set of real $n \times n$ matrices with one complex conjugate pair of eigenvalues

of algebraic multiplicity two and geometric multiplicity one and all other eigenvalues simple. Define $X = X_r \cup X_i$. From Theorem 2, the generic case of a one-dimensional $\mathcal{N}(G_u^0)$ implies that $A^0 \in X$. We now show that X_r is generic in X.

For each $A \in X_r$, we associate $V(A) \equiv (A, Y, \lambda, d_3, \ldots, d_n)$, where Y is a real $n \times n$ matrix, λ is the unique multiple eigenvalue of A, and $d_3, \ldots, d_n$ are real numbers. In the case in which all the eigenvalues of A are real, the columns of Y can be considered as the generalized eigenvectors of A and d_j as the eigenvalues. If A has a complex eigenvalue μ with eigenvector z, then we could take $\mu = d_3 + id_4$ and $z = y_3 + iy_4$, for example. (y_j denotes the jth column of Y.) Note that $(\overline{\mu}, \overline{z})$ is also an eigenvalue–eigenvector pair of A. The point is that the information contained in Y and d_j is enough to determine the eigenvalues and eigenvectors of A. In the case in which all eigenvalues are real, $V(A)$ must satisfy

$$AY = YJ, \quad J = \left[\begin{array}{cc|ccc} \lambda & 1 & & & \\ 0 & \lambda & & & \\ \hline & & d_3 & & \\ & & & \ddots & \\ & & & & d_n \end{array}\right].$$

If complex eigenvalues exist, the above must be appropriately modified. In addition, there are n normalization equations for the eigenvectors. Thus, $V(A)$ consists of $2n^2 + n - 1$ real variables which must satisfy $n^2 + n$ real polynomial equations and thus has $n^2 - 1$ degrees of freedom.[1]

For $A \in X_i$, let $V(A) \equiv (A, Y, \lambda_r, \lambda_i, d_5, \ldots, d_n)$, where $\lambda \equiv \lambda_r + i\lambda_i$ is the complex eigenvalue of A of algebraic multiplicity two. The Jordan form (in the case in which all other eigenvalues are real) is

$$J = \left[\begin{array}{cc|cc|ccc} \lambda & 1 & & & & & \\ 0 & \lambda & & & & & \\ \hline & & \overline{\lambda} & 1 & & & \\ & & 0 & \overline{\lambda} & & & \\ \hline & & & & d_5 & & \\ & & & & & \ddots & \\ & & & & & & d_n \end{array}\right].$$

Thus, $V(A)$ consists of $2n^2 + n - 2$ real variables and must also satisfy $n^2 + n$ real equations and thus it has $n^2 - 2$ degrees of freedom. Hence we see that X_r is generic.

We remark that the equations $AY = YJ$ and the normalization equations are linearly independent. If one normalization equation is omitted, then the length of some eigenvector is not uniquely determined. Also, if one of the real equations in $AY = YJ$ is omitted, then we may not have an eigenvalue–eigenvector pair. Also, in the above calculation we actually include matrices with eigenvalues of higher multiplicities and other multiple eigenvalues (besides λ). This is acceptable because they are nongeneric in X. □

At simple quadratic folds and simple quadratic bifurcation points the eigenvalue has algebraic multiplicity two and geometric multiplicity one. At both cubic fold and simple pitchfork bifurcation points the algebraic and geometric multiplicities are

[1] In the language of algebraic geometry, $V(A)$ is a variety and the degrees of freedom correspond to the dimension of the variety.

TABLE 1

Summary of some of the different types of points at a singular Jacobian G_u^0. With the exception of the quadratic fold, additional generic conditions must be satisfied for all.

	$\psi^* G_t^0 \neq 0$	$\psi^* G_t^0 = 0$
$\psi^* G_{uu}^0 \phi^2 \neq 0$	simple quadratic fold	simple quadratic bifurcation
$\psi^* G_{uu}^0 \phi^2 = 0$	simple cubic fold	simple pitchfork bifurcation

three and one, respectively. See Table 1. The Jacobian G_u^0 of course may have other types of nongeneric singularities. For example, the eigenvalue may have multiplicities three and two, respectively. However, these are nongeneric and unlikely to occur in practice.

The significance of the above theory is that in practice we encounter only simple real quadratic folds, and this is the route by which real eigenpaths become complex.

2.7. A bound on the number of bifurcation points. It is not difficult to show that at a real or complex bifurcation point of (2) the algebraic multiplicity of the eigenvalue of $A(t)$ is at least two. Let

$$p(t, \lambda) \equiv \det(A(t) - \lambda I).$$

Since $A(t)$ is linear in t, the above is a polynomial in (t, λ) of degree n. In fact, p can be written in the form

$$p(t, \lambda) = a_0(t) + a_1(t)\lambda + \cdots + a_n(t)\lambda^n, \tag{14}$$

where $a_i(t)$ is a polynomial in t of degree at most $n - i$ for $i = 0, \ldots, n$ and $a_n(t) = (-1)^n$. Define

$$q(t, \lambda) = \frac{\partial p(t, \lambda)}{\partial \lambda}.$$

From (14), it is easy to show that q is a polynomial of degree $n - 1$. At a bifurcation point (t, λ) we must have

$$p(t, \lambda) = q(t, \lambda) = 0.$$

This is a system of two polynomial equations of degrees n and $n - 1$ in two variables. By Bézout's theorem, it has at most $n(n - 1)$ roots. *Hence the eigenpaths collectively can have at most $n(n - 1)$ bifurcation points.*

We remark that some of these roots may have a complex time t and that some roots may lie outside the region of interest (i.e., $t \in [0, 1]$). In practice we usually see on the order of n bifurcation points.

3. Numerical algorithm. In this section, we describe the numerical implementation of the homotopy algorithm including choice of the initial matrix A_0, stepsize selection, and transition from real to complex eigenpairs and vice versa. For a more thorough treatment of some of these topics, see Keller [14] and Allgower and Georg [1].

Suppose that we have computed the eigenpairs at time t_0. The normalization equation for the eigenvector x at the new time is taken to be

$$x_0^* x - 1 = 0,$$

where x_0 is the eigenvector at time t_0. We always perform real arithmetic so that the pseudoarclength formulation (3) is written as an equivalent system of $2n + 4$ real equations whenever we are following a complex eigenpath.

3.1. Choice of initial matrix A_0. The constraint that the eigenpairs of A_0 be computable quickly severely limits the choice of A_0. Ideally, A_0 should be chosen so that the number of real and complex bifurcation points are minimized. This is because there is extra work involved in locating real fold points. In the example shown in Figure 2, A_0 is a diagonal matrix. By simply reordering the diagonal elements of this A_0 it is possible for the eigenpaths to have just three real fold points. This is the minimum possible because this A_1 has six complex eigenvalues. There are no "unnecessary" fold points. Another desirable property of A_0 is that the eigenpaths be well separated. This decreases the chance of the path-jumping phenomenon. However, it seems extremely difficult to choose a priori an initial matrix which has all of the above properties.

We tried three different kinds of initial matrices: real diagonal, real block diagonal with 2×2 diagonal blocks, and block upper triangular with 2×2 diagonal blocks. We now describe them in more detail.

The real diagonal initial matrix is defined as follows. Let a denote the trace of A_1 divided by n, the size of the matrix. This is the average value of the eigenvalues of A_1. Let ρ be the square root of the maximum of the Gerschgorin radii of A_1. Define the diagonal elements of A_0 as equally distributed points in $[a-\rho, a+\rho]$ in ascending order. There is no theoretical justification for this choice of A_0 except that the eigenvalues are initially simple and the eigenvectors are just the standard basis vectors. Without the square root in the definition of ρ, numerical experiments on random matrices show that the initial eigenvalue distribution is too spread out. An alternative is to simply use the diagonal part of A_1 as the initial matrix. One problem here is that this initial matrix may have multiple eigenvalues, leading to potential difficulties.

For a real diagonal initial matrix, the eigenpaths are real initially. As we shall see, the resultant homotopy usually has a large number of "unnecessary" fold points. As an attempt to remedy the situation, we tried initial matrices which have complex eigenvalues. One avenue is to try an A_0 which is real block diagonal with 2×2 diagonal blocks of the form

$$\begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix}.$$

The eigenvalues of this block are $\alpha \pm i\beta$. The pairs (α, β) are chosen as uniformly distributed in the square box in the complex plane with center at the point $a + 0i$ (the average of the eigenvalues of A_1) and width 2ρ, where ρ was defined in the above paragraph. Now the eigenpaths start out complex. Since the complex space is much bigger than the real space, there is less likelihood of two eigenpaths venturing close together (hence less chance of path jumping) and less possibility of encountering fold points.

The final kind of initial matrix we consider is block upper triangular with 2×2 diagonal blocks. The upper triangular part of the matrix is taken to be the upper

triangular part of A_1 and the 2×2 diagonal blocks are as defined above. We define the 2×2 diagonal blocks this way, instead of copying those of A_1, to avoid possible multiple eigenvalues in the beginning. The eigenpairs of this initial matrix can be found quickly. The motivation for this initial matrix is that it is closer to A_1 than previous initial matrices. A smaller $\|A_1 - A_0\|$ should lead to straighter eigenpaths and possibly fewer fold points.

Some very limited experiments with 100×100 random matrices confirm our observations. A diagonal initial matrix leads to many more fold points than the other two initial matrices. The third type of initial matrix performs marginally better than the second type.

3.2. Transition at real fold points. We first describe the transition from a real eigenpath to a complex one. When it detects that it is going backwards in time, then, generically, a real fold has been passed. By the theory of the last section, there must be a complex conjugate pair of solutions on the opposite side of the real fold. We first get a more accurate location of the fold point by using the secant method to approximate the point at which $dt/ds = 0$. (Recall that this is a necessary condition at a fold point.) With the augmented system, the Jacobian (4) is nonsingular, so there is no numerical difficulty in the task. We store the location of this fold point in a table for later reference. Using the tangent vector ϕ at the fold point, we solve problem (2) in complex space at a later time. This is done by carrying out the Euler–Newton continuation with the initial tangent $i\phi$, in accordance with the theory of Henderson and Keller.

When the partner of the above path comes from the other arm of the same fold, it checks that the fold point has been visited before and stops further computation. This way, only one path of a complex conjugate pair of eigenpaths is computed.

The reverse of the above situation also arises, although less frequently. That is, time decreases while advancing along a complex path. Generically, there must be a real fold on the opposite side of this complex path. Once the fold point has been located, we compute the real tangent vector ϕ. We then apply the Euler–Newton continuation in both the directions ϕ and $-\phi$. See Figure 7. Because the problem is being solved in real space, there is no chance of converging back to the complex solution. On a parallel computer, a node which became idle at another fold point can be invoked to carry out the computation along one of these directions. If we begin with k complex eigenpaths, we may end up with many more than k eigenpaths because of these complex-to-real bifurcations. Fortunately, in practice, at most a few more have been encountered.

3.3. Computing the tangent. Suppose two eigenpairs u_0 and u_1 have been found. We wish to compute the tangent vector at t_1. In formulation (3), we have

$$F_u^1 \dot{u}_1 + F_t^1 \dot{t}_1 = 0,$$

where the superscript ^{1} denotes the evaluation of the Jacobian at (u_1, t_1) and the dot denotes the s derivative. For a unit tangent, we require in addition that

$$\dot{u}_1^* \dot{u}_1 + \dot{t}_1^2 = 1. \tag{15}$$

Note that the above two equations define the tangent up to a sign. To ensure that we are always computing in the same direction, we further impose the condition

$$\Re(\dot{u}_0^* \dot{u}_1) + \dot{t}_0 \dot{t}_1 > 0.$$

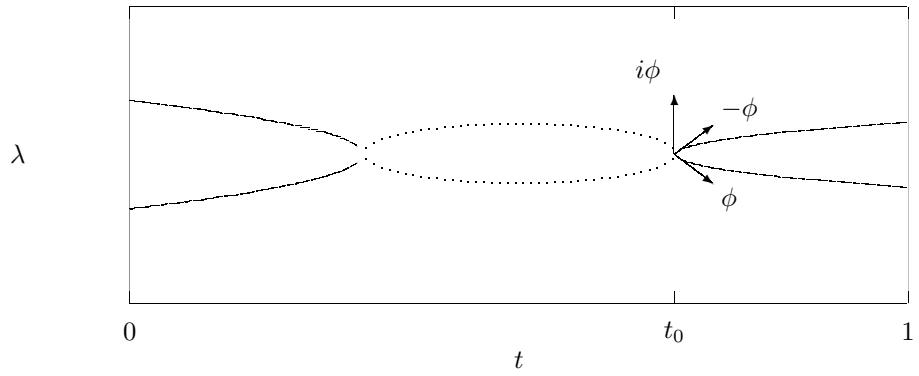


FIG. 7. *Transition from a complex solution to a real solution at a fold point. Dotted lines denote complex solutions.*

Because (15) is nonlinear, we instead solve the linear system (when u_0 is real)

$$\begin{bmatrix} F_u^1 & F_t^1 \\ \dot{u}_0^T & \dot{t}_0 \end{bmatrix} \begin{bmatrix} \phi \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The tangent $(\dot{u}_1, \dot{t}_1)$ is obtained by normalizing the solution of the above system.

3.4. Selection of stepsize. Suppose we have the two eigenpairs u_0 and u_1. We obtain stepsize δs_2 for u_2 as follows:

$$\delta s_2 = \delta s_1 \ (\Re(\dot{u}_0^* \dot{u}_1) + \dot{t}_0 \dot{t}_1 + .5),$$

where δs_1 is the stepsize used to obtain u_1. The idea is that when the two previous tangents are parallel we increase the stepsize by 50%. If the tangents are perpendicular, we decrease the stepsize by a half. We use the above scheme until the time is close to one, at which time we solve the system $G(u, 1) = 0$.

Whenever a Newton iteration fails to converge after, say, six iterations, we restart it with a stepsize that is one-half of the original one.

3.5. Path jumping. Path jumping is a serious problem for the homotopy method. This is the phenomenon in which the Newton iteration converges to another eigenpath. This occurs when the stepsize is overly ambitious or the linear system involved in the solution of a Newton iterate has a large condition number. The latter situation arises whenever eigenvalues are poorly separated.

An elegant method of detecting path jumping is available when the matrix is symmetric tridiagonal with nonzero off-diagonal elements (Li and Rhee [17]). They employ the Sturm sequence property of symmetric matrices. Li and Zeng [19] can also detect path jumping in the case in which the eigenpath is real and the matrix is in Hessenberg form. However, no satisfactory procedure is known for general matrices. One inefficient way is to use the property that the sum of the eigenvalues of a matrix is equal to the trace of the matrix. Noting that

$$\mathrm{Tr}(A(t)) = \mathrm{Tr}(A_0) + t(\mathrm{Tr}(A_1 - A_0)),$$

almost all path jumps can be detected by comparing the sum of the computed eigenvalues and the above expression for the trace of $A(t)$. However, this does not tell us

which path has jumped, and hence it is necessary to recompute the last step for all eigenpaths. Other drawbacks include the necessity of synchronizing the computation of the eigenpaths and the fact that this method works only if all the eigenpaths are computed.

Our approach is perhaps the simplest, but certainly not the best. We keep track of the initial eigenvalue (at $t = 0$) of each eigenpath, and for each eigenpath that has been computed more than once (this is checked at $t = 1$) we repeat the entire calculation for those eigenpaths with a smaller stepsize.

3.6. Parallel aspects. The homotopy method is fully parallel because each eigenpath can be computed independently of the others. If the sparse matrix $A(t)$ can be stored in each node, then there is no communication overhead at all other than the trivial broadcast of the location of fold points.

3.7. Homotopy algorithm of Li, Zeng, and Cong. Li, Zeng, and Cong [20] use a different strategy in their homotopy algorithm. They first use Householder transformations to reduce the given matrix to a similar matrix A_1 in upper Hessenberg form. Their initial matrix A_0 is the same as A_1 except that one subdiagonal entry is set to zero. They use a divide-and-conquer strategy to obtain the eigenpairs of A_0. Because A_0 is very close to A_1, the eigenpaths will be nearly straight and path jumping is much less of a problem here. The performance of this method is very encouraging. However, it requires storage of the entire matrix plus large amounts of work storage. For another approach to finding the eigenvalues using homotopy, see Lenard [15].

4. Numerical results. We have done very limited testing on random matrices and matrices arising from the finite difference approximations of partial differential equations. The tests were performed on SUN Sparc workstations. In our code, we computed the eigenpairs one at a time. As mentioned already, in a parallel code each eigenpair can be assigned to a separate processor.

We use an initial $ds = .1$, a final tolerance of 10^{-12}, and an intermediate tolerance of 10^{-4}. The final tolerance means that the stopping criterion for the Newton iteration is that the norm of the Newton step is less than 10^{-12} at $t = 1$. Intermediate tolerance refers to the stopping criterion at $t < 1$. The criterion for stopping the iteration to locate a fold point is $|\dot{t}| < 10^{-3}$.

Empirically, we notice that the eigenpaths move in a relatively simple fashion as t progresses. That is, there are no wild oscillations. Thus, the homotopy method has the potential to find efficiently a few special eigenvalues, for example, those with the largest real part. Such eigenvalues are of interest in linear stability theory for partial differential equations. It is in this area that we believe the homotopy method will be most useful.

Our first set of test examples comes from the usual second-order finite difference discretization of the elliptic operator

$$\triangle + f(x,y)\frac{\partial}{\partial x} + g(x,y)\frac{\partial}{\partial y} + p(x,y) \tag{16}$$

on a rectangle of size 1×1.2 with homogeneous Dirichlet boundary condition. We choose the initial matrix as the discrete Laplacian whose eigenpairs are known. We make the following changes to the algorithm to account for the nature of the problem. Assuming a uniform mesh size h in both x and y, the modified equation for the tangent

TABLE 2
Execution times for five eigenpaths of matrices of various sizes corresponding to the discretizations of a PDE with different grid sizes.

size	238	696	1394	3510	10622
time (sec)	4	18	49	197	1619

is

$$h^2\dot{v}_0^*\dot{v} + \dot{\lambda}_0\dot{\lambda} + \dot{t}_0\dot{t} = 1.$$

Here, $\dot{v}$ denotes the s derivative of the eigenvector and the subscript denotes the corresponding quantity at the previous time t_0. The reason for the modification is that this approximates the underlying continuous equation

$$\int \dot{v}_0\dot{v}dxdy + \dot{\lambda}_0\dot{\lambda} + \dot{t}_0\dot{t} = 1.$$

Similarly, we employ the following pseudoarclength condition:

$$h^2\dot{v}_0^*(v - v_0) + \dot{\lambda}_0(\lambda - \lambda_0) + \dot{t}_0(t - t_0) - ds = 0.$$

For the numerical experiments, we take a uniform 95×114 grid leading to a matrix of dimension 10622. We follow the five paths whose initial eigenvalues are largest with the aim of computing the five eigenvalues of the PDE having the largest real parts. Our Fortran code uses GMRES [27] to solve each linear system. (An alternative is the QMR method of Freund and Nachtigal [8].) Here, only the nonzero entries of the matrices need be stored. The average number of time steps per eigenpath is 5 and the number of Newton iterations per step is 1. The program successfully computed the five eigenpairs with the five largest real parts in a number of examples that we tried. These computed paths all turned out to be real. Execution times for various choices of the coefficients of the PDE are between 27 and 28 minutes.

In Table 2, we give the execution times for computing five eigenpaths for the PDE with coefficients $f = e^x - 2y^2, g = y^2\cos(2x), p = 0$ for various grid sizes. The maximum dimension of the Krylov subspace, a parameter of GMRES, was set at 100 for all the test runs. Hence, the execution time for smaller matrices is more favorable than for larger matrices. The complexity is slightly less than $O(n^2)$.

We also tried a symmetric problem (with $f = g = 0$ and various choices of p). The execution times are between 16 and 22 minutes for matrices of size 10622.

We have not been able to devise a mechanism to guarantee that an eigenpath will end up (at $t = 1$) having an eigenvalue with the largest real part, even for the scalar PDE above. Problems which arise in practice (for example, in fluid mechanics) often involve systems of PDEs. It would be very difficult to obtain any theoretical result in this direction.

As our final illustration, we compute singular points of a parameter-dependent scalar PDE which arises in population biology. The PDE is

$$\triangle u + \alpha f(u) + \gamma u_x = 0$$

on a rectangular domain of sides of widths 1 and 1.2, and homogeneous Dirichlet boundary conditions are imposed. This is a population model for insects in a domain with a constant prevailing wind of strength γ in the $-x$ direction. Here, u represents

the population of the insects and α is a parameter depending on the birth rate and diffusion coefficient. The boundary conditions mean that the exterior of the domain is completely hostile to the insects. See Murray [23] for further details. We will only consider the Fisher model; i.e., $f(u) = u(1-u)$. The problem is to determine values of α for which the PDE becomes singular. Such points are of interest because bifurcation typically occurs there. These points are special because there solutions lose/gain stability. Singular points occur when the corresponding linearized eigenvalue problem (linearized about $u = 0$)

$$\triangle v + \alpha f'(0)v + \gamma v_x = \lambda v$$

has a zero eigenvalue. Hence the problem reduces to finding a zero eigenvalue of the matrix which arises from the discretization of the above equation.

Here is how the algorithm proceeds. Using the matrix which arises from the discretization of the Laplacian as the initial matrix, we use the homotopy algorithm to find the largest eigenvalue at $\alpha = 0$ (where all the eigenvalues are negative). We then follow this eigenpath at increasing values of α until the eigenvalue becomes positive. At that point, we use the secant method to locate the zero of the eigenvalue (as a function of α). For the eigenvalue problem at α_{i+1}, we use the corresponding matrix at α_i as the initial matrix.

Dividing the rectangle into a uniform 95×114 grid, we obtain a matrix of size 10622. For a wind strength $\gamma = 1$, the code computed the eigenvalue at $\alpha = 0, 5, 10, 15,$ and 20. Discovering that the eigenvalue becomes positive at the last value of α, it proceeded to compute the critical value $\alpha^* = 16.97\ldots$ in one step of the secant method. It found α^* with the eigenvalue at that point on the order of 10^{-12}. The entire procedure took 534 seconds, with the first eigenvalue solve at $\alpha = 0$ taking 416 seconds and the rest of the calculation taking about 120 seconds. This example illustrates the power of the homotopy method. When the initial matrix and the final matrix do not differ significantly, the eigenvalues can be found quite rapidly.

We have also tried the Lanczos code of Freund, Gutknecht, and Nachtigal [7] on problem (16) with a matrix of size 10622. With 500 Lanczos iterations, it computed the same five eigenpairs in about 280 seconds for each problem. This code is superior to our code in terms of both efficiency and robustness. However, it suffers the same problem as ours in that it cannot guarantee which eigenvalues it computed.

5. Conclusion. We have presented a homotopy method for computing the eigenpairs of a real matrix. Starting with a matrix with known eigenpairs, Euler–Newton continuation is used to advance the eigenpaths. A real eigenpath will remain real unless it encounters a real fold point. On the opposite side of this fold point, two complex conjugate eigenpairs emerge. The reverse situation in which two complex conjugate eigenpairs meeting at a real fold point with two real paths bifurcating to the right also occurs. By restricting the solutions in the real space, we have shown how to deal with these transitions without numerical difficulties.

The storage requirement is on the order of the number of nonzero elements of the matrix, and thus it is attractive for computing a few eigenpairs of a large, sparse matrix. This together with the fully parallel nature of the algorithm may make it a competitive method for the large, sparse nonsymmetric eigenvalue problem. However, several formidable obstacles must first be overcome. The path-jumping problem has already been mentioned. Another is the absence of a robust general-purpose iterative linear solver. GMRES had considerable convergence difficulties for general matrices. Even for the PDE examples that we tried, it encountered convergence problems when

computing interior eigenvalues. The homotopy method also has difficulty whenever eigenvalues are clustered together. This occurs even if the eigenvectors are orthonormal. The difficulty lies in the fact that eigenvectors cannot be computed accurately by a straightforward application of the inverse iteration (or Newton's method) if the corresponding eigenvalues are clustered together. One solution is to compute the clustered eigenvalues by subspace iteration. However, if the initial matrix is not well chosen, then it is possible that eigenvalues which are far apart initially at $t = 0$ drift together at some point $t \leq 1$. Choosing a good initial matrix for the homotopy which would minimize the number of bifurcation points and keep the eigenpaths well separated is another open problem. Finally, we would like to determine selected eigenvalues (for example, those with the largest real part) by following just one or two eigenpaths. The homotopy method seems to be a very efficient method for locating singular points of bifurcation problems.

The history of the homotopy method as a computational tool for the eigenvalue problem is rather short. We hope this work will stimulate further interest in this area.

Acknowledgment. We thank the referees for suggesting numerous improvements to the original draft.

REFERENCES

[1] E. L. Allgower and K. Georg, *Numerical Continuation Methods*, Springer-Verlag, New York, Berlin, 1990.

[2] Z. Bai, D. Day, and Q. Ye, *ABLE: An Adaptive Block Lanczos Method for Non-Hermitian Eigenvalue Problems*, Technical report, University of Kentucky, Lexington, KY, 1995.

[3] Z. Bai and J. Demmel, *On a block implementation of Hessenberg multishift QR iteration*, Internat. J. High Speed Comput., 1 (1989), pp. 97–112.

[4] Z. Bai and J. Demmel, *Design of a Parallel Nonsymmetric Eigenroutine Toolbox, Part* I, Technical report UCB/CSD-92-718, U. C. Berkeley, Berkeley, CA, 1992.

[5] J. Cullum and R. A. Willoughby, *A practical procedure for computing eigenvalues of large sparse nonsymmetric matrices*, in Large-Scale Eigenvalue Problems, Math. Stud. 127, J. Cullum and R. A. Willoughby, eds., North–Holland, Amsterdam, 1986.

[6] J. J. Dongarra and M. Sidani, *A parallel algorithm for the nonsymmetric eigenvalue problem*, SIAM J. Sci. Comput., 14 (1993), pp. 542–569.

[7] R. W. Freund, M. H. Gutknecht, and N. M. Nachtigal, *An implementation of the look-ahead Lanczos algorithm for non–Hermitian matrices*, SIAM J. Sci. Comput., 14 (1993), pp. 137–158.

[8] R. W. Freund and N. M. Nachtigal, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.

[9] G. Golub and C. Van Loan, *Matrix Computations*, 2nd ed., The John Hopkins University Press, Baltimore, MD, 1989.

[10] M. E. Henderson, *Complex Bifurcation*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 1985.

[11] M. E. Henderson and H. B. Keller, *Complex bifurcation from real paths*, SIAM J. Appl. Math., 50 (1990), pp. 460–482.

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

[13] H. B. Keller, *Numerical solutions of bifurcation and nonlinear eigenvalue problems*, in Applications of Bifurcation Theory, P. H. Rabinowitz, ed., Academic Press, New York, 1977.

[14] H. B. Keller, *Lectures on Numerical Methods in Bifurcation Problems*, Springer-Verlag, Berlin, 1987.

[15] C. T. Lenard, *A Homotopy Method for Eigenproblems*, Technical report, Australia National University, Canberra, 1990.

[16] K. Li and T. Y. Li, *An algorithm for symmetric tridiagonal eigenproblems: Divide and conquer with homotopy continuation*, SIAM J. Sci. Comput., 14 (1993), pp. 735–751.

[17] T. Y. Li and N. H. Rhee, *Homotopy algorithm for symmetric eigenvalue problems*, Numer. Math., 55 (1989), pp. 265–280.

[18] T. Y. Li and X. Wang, *Higher order turning points*, Appl. Math. & Comp., 64 (1994), pp. 155–166.
[19] T. Y. Li and Z. Zeng, *Homotopy-determinant algorithm for solving nonsymmetric eigenvalue problems*, Math. of Computation, 59 (1992), pp. 483–502.
[20] T. Y. Li, Z. Zeng, and L. Cong, *Solving eigenvalue problems of real nonsymmetric matrices with real homotopies*, SIAM J. Numer. Anal., 29 (1992), pp. 229–248.
[21] T. Y. Li, H. Zhang, and X. H. Sun, *Parallel homotopy algorithm for the symmetric tridiagonal eigenvalue problem*, SIAM J. Sci. Comput., 12 (1991), pp. 469–487.
[22] S. H. Lui, *Part* II: *Parallel Homotopy Method for the Real Nonsymmetric Eigenvalue Problem*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 1992.
[23] J. D. Murray, *Mathematical Biology*, Springer-Verlag, New York, Berlin, 1989.
[24] A. Ruhe, *The rational Krylov algorithm for nonsymmetric eigenvalue problems.* III: *Complex shifts for real matrices*, BIT, 34 (1994), pp. 165–176.
[25] Y. Saad, *Numerical solution of large nonsymmetric eigenvalue problems*, Computer Physics Comm., 53 (1989), pp. 71–90.
[26] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, 1992.
[27] Y. Saad and M. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.
[28] G. Shroff, *A parallel algorithm for the eigenvalue and eigenvectors of a general complex matrix*, Numer. Math., 58 (1991), pp. 779–805.
[29] D. C. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385.
[30] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.
[31] Z. H. Yang and H. B. Keller, *A direct method for computing higher order folds*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 351–361.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, pp. 334–347, April 1997

006

NORMS AND INEQUALITIES RELATED TO SCHUR PRODUCTS OF RECTANGULAR MATRICES*

WENCHAO HUANG†, CHI-KWONG LI‡, AND HANS SCHNEIDER†

Abstract. We consider operator norms on rectangular matrices. When the underlying vector norms are semiabsolute/absolute and the matrices are nonnegative, we obtain an inequality involving Schur (Hadamard) products of fractional Schur powers of matrices and the product of the fractional powers of the norms of the matrices. This leads naturally to the concept of fractional Schur (nonnegative) submultiplicativity factors for a norm. As a corollary, we obtain a necessary and sufficient condition for a norm to be Schur submultiplicative on nonnegative matrices. We also consider the relation of the least fractional Schur submultiplicativity factor and the least Schur submultiplicativity factor for general matrices, and we prove some necessary and sufficient conditions for Schur submultiplicativity.

Key words. matrix, Schur (Hadamard) product, submultiplicativity, norm

AMS subject classifications. 15A60, 15A45

PII. S0895479896298877

1. Introduction.

Let $\mathbf{F}$ be $\mathbf{R}$ or $\mathbf{C}$, where $\mathbf{R}$ and $\mathbf{C}$ are the fields of real numbers and complex numbers, respectively. Denote by $\mathbf{F}^n$ the vector spaces of all n-dimensional column vectors over $\mathbf{F}$, and denote by $\mathbf{F}^{mn}$ the set of all $m \times n$ matrices over $\mathbf{F}$. A matrix $A \in \mathbf{F}^{mn}$ is called *nonnegative* if A is nonnegative entrywise.

The *Schur* (*Hadamard*) *product* of two matrices $A = (a_{ij})$ and $B = (b_{ij})$ in $\mathbf{F}^{mn}$, denoted by $A \circ B$, is defined as $A \circ B = (c_{ij})$, where $c_{ij} = a_{ij}b_{ij}$.

Let μ and ν be norms on the vector spaces $\mathbf{F}^m$ and $\mathbf{F}^n$, respectively. Let $[\mu/\nu]$ be the operator norm on $\mathbf{F}^{mn}$ induced by the norms μ and ν; i.e.,

$$[\mu/\nu](A) := \max\{\mu(Ax) : \nu(x) = 1\} \tag{1}$$

for all $A \in \mathbf{F}^{mn}$. In this paper, we focus on the behavior of the *operator norms* $[\mu/\nu]$ under *Schur multiplication* of rectangular matrices. For further definitions of terms italicized below see section 2 and the beginnings of the following sections.

In the remainder of this introduction, we shall assume that μ is a *semiabsolute* norm and that ν is an *absolute* norm.

In section 3, we investigate the behavior of norms and operator norms on Schur products of fractional Schur powers of nonnegative matrices. We show that the least *fractional Schur submultiplicativity factor* for the norm $[\mu/\nu]$ is determined by the values of the norms μ and ν on the unit vectors, and that it is equal to the least *nonnegative submultiplicativity factor* for the same norm; see Proposition 3.3, Theorem 3.5, and Remark 3.6(a). Our proof uses a generalization of Hölder's inequality [Ho1889], which was shown in [JN91]. In Corollary 3.7, we give two necessary and sufficient conditions for $[\mu/\nu]$ to be *nonnegative Schur submultiplicative*.

In section 4, we investigate the behavior of the norms and operator norms on Schur products of (general) vectors and matrices, and this leads naturally to the study of

* Received by the editors February 15, 1996; accepted for publication by R. A. Horn May 1, 1996. http://www.siam.org/journals/simax/18-2/29887.html

† Department of Mathematics, University of Wisconsin–Madison, Madison, WI 53706 (whuang@math.wisc.edu, hans@math.wisc.edu). The research of these authors was supported in part by NSF grants DMS–9123318 and DMS–9424346.

‡ Department of Mathematics, The College of William and Mary, Williamsburg, VA 23187 (ckli@cs.wm.edu). The research of this author was supported in part by a NATO grant.

Schur submultiplicativity factors. Example 4.6 (due to Bit-Shun Tam) shows that the least Schur submultiplicativity factor may be larger than the least nonnegative Schur submultiplicativity factor. We prove some inequalities that must hold for these two types of factors; see Theorem 4.4 and Remark 4.5. In Corollary 4.8, under the more restrictive assumption that either μ or ν^D is a weighted l_∞ norm, we give a necessary and sufficient condition for an operator norm $[\mu/\nu]$ to be *Schur submultiplicative*. We end section 4 with two open questions.

In section 5, we introduce the concept of the *Schur K-operator norm* induced by a norm. If K is the cone of all nonnegative matrices in $\mathbf{F}^{mn}$, we apply a result on the limits of norms of Schur powers of nonnegative vectors to show that, for any nonnegative matrix, the Schur K-operator norm induced by $[\mu/\nu]$ equals the elementwise l_∞ norm; see Corollary 5.3.

At the beginning of each of the following sections, a summary of the results in that section and some useful definitions will be given.

The subject of norms on Schur products was investigated in [S11], [Ok87], [MKS84], [On84], [W86], [HJ87], [AHJ87], [Zh88], [N89], [HM90], [JN91], and [AG94], etc. (also see [H90], [HJ91, Chapter 5], and some further references at the beginning of section 4). The behavior of (subadditive) norms and corresponding (sub)multiplicativity factors on more general algebras was studied previously; see, e.g., [G90], [AG90], [AGL92], [AGL93a], and [AGL93b]. Fractional Schur powers of nonnegative matrices were investigated in [FH77], [KO85], [EJS88], and [MP96].

2. Preliminaries. A *norm* $\mathcal{N}$ on $\mathbf{F}^{mn}$ is a function $\mathcal{N} : \mathbf{F}^{mn} \mapsto \mathbf{F}$ satisfying the following axioms for all $A, B \in \mathbf{F}^{mn}$:

(1) $\mathcal{N}(A) \geq 0$, and $\mathcal{N}(A) = 0$ if and only if $A = 0$;
(2) $\mathcal{N}(cA) = |c|\mathcal{N}(A)$ for all $c \in \mathbf{F}$;
(3) $\mathcal{N}(A + B) \leq \mathcal{N}(A) + \mathcal{N}(B)$.

Next we state some known results on norms which are used in what follows. Recall that the *dual norm* μ^D of any norm μ on $\mathbf{F}^n$ with the standard inner product is defined by (see [BSW61] or [HJ85, 5.4.12])

$$\mu^D(x) := \max\{|x^*y| : \mu(y) = 1\} \tag{2}$$

for all $x \in \mathbf{F}^n$. Using the concept of dual norm, one has a simple formula for the operator norm $[\mu/\nu]$, as shown in the following proposition (see [Ba63] for the first formula for $[\mu/\nu]$; the second formula can be obtained immediately by (1) and (2)).

PROPOSITION 2.1. *For any $A \in \mathbf{F}^{mn}$, $x \in \mathbf{F}^m$, and $y \in \mathbf{F}^n$ we have*

$$[\mu/\nu](A) = \max\{|x^*Ay| : \mu^D(x) = 1, \nu(y) = 1\}$$

and

$$[\mu/\nu](xy^*) = \mu(x)\nu^D(y).$$

Let μ be a norm on $\mathbf{F}^m$. Then μ is called *absolute* if $\mu(|x|) = \mu(x)$ for all $x \in \mathbf{F}^m$, where $|x|$ is the vector obtained from x by replacing the entries of x by their magnitudes. The norm μ is called *semiabsolute* if $\mu(x) \leq \mu(|x|)$ for all $x \in \mathbf{F}^m$. Further, μ is called *quasi-monotonic* if $\mu(x) \leq \mu(y)$ whenever $0 \leq x \leq y$, where $x \geq 0$ means that x is entrywise nonnegative and $x \leq y$ means that $y - x \geq 0$.

It is well known that a norm μ is absolute if and only if it is a *monotonic* norm in the sense that $\mu(x) \leq \mu(y)$ whenever $|x| \leq |y|$ (see [BSW61], [SS75], [G89], and [HJ85, 5.5.10]). By [HM90, Lemma 2.4] (cf. [JN91, Theorem 2 (D$\longrightarrow$B)]), it can easily be

shown that a norm μ is semiabsolute if and only if it is a *semimonotonic* norm in the sense that $\mu(x) \leq \mu(y)$ whenever $|x| \leq y$. The definitions of semimonotonic norms and quasi-monotonic norms can be found in [G90]. Obviously, every monotonic norm is semimonotonic, and every semimonotonic norm is quasi-monotonic (see [G90]).

A norm μ on $\mathbf{C}^m$ is called *conjugate invariant* if $\mu(\bar{x}) = \mu(x)$ for all $x \in \mathbf{C}^m$, where $\bar{x}$ is the vector obtained from x by changing its entries to their conjugates. Clearly, any absolute norm is conjugate invariant.

A norm μ on $\mathbf{F}^m$ is called *axis standardized* if $\mu(e_i^{(m)}) = \mu(e_j^{(m)})$ for all $1 \leq i, j \leq m$, where $e_i^{(m)}$ denotes the ith unit vector in $\mathbf{F}^m$.

Suppose μ and ν are norms on $\mathbf{F}^m$. We say that ν is a *weighted μ norm* if there exists an entrywise positive vector $w \in \mathbf{F}^m$ such that $\nu(x) = \mu(w \circ x)$ for all $x \in \mathbf{F}^m$.

For $A = (a_{ij}) \in \mathbf{F}^{mn}$, denote by $|A|$ the matrix with $|a_{ij}|$ as its (i, j) element.

PROPOSITION 2.2. *Let μ be a semiabsolute norm on $\mathbf{F}^m$ and let ν be an absolute norm on $\mathbf{F}^n$. Then, for all $A \geq 0$ in $\mathbf{F}^{mn}$, we have*

$$[\mu/\nu](A) = \max\{\mu(Ax) : \nu(x) = 1, x \geq 0\}.$$

Further, for all $B \in \mathbf{F}^{mn}$, we have

$$[\mu/\nu](B) \leq [\mu/\nu](|B|). \tag{3}$$

Proof. Assume $x \in \mathbf{F}^n$ and $\nu(x) = 1$. Since μ is semiabsolute, we have

$$\mu(Ax) \leq \mu(|Ax|) \leq \mu(A|x|).$$

Since ν is absolute, it follows that $\nu(|x|) = 1$, and the above inequalities show that $[\mu/\nu](A) \leq \max\{\mu(Ax) : \nu(x) = 1, x \geq 0\}$. Thus the first formula follows.

The second assertion follows immediately from the first formula. ∎

Remark 2.3.

(a) We observe that $[\mu/\nu]$ is also conjugate invariant if μ and ν are conjugate invariant.

(b) By Proposition 2.2, the operator norm $[\mu/\nu]$ induced by a semiabsolute norm μ and an absolute norm ν is semiabsolute.

Since $\nu(e_i^{(n)})\nu^D(e_i^{(n)}) = 1$ for all i if ν is an absolute norm on $\mathbf{F}^n$, we can deduce the following result concerning $[\mu/\nu](E_{ij})$ from Proposition 2.1, where E_{ij} denotes the standard unit matrix with the (i, j) element equal to 1 and the other elements equal to 0.

COROLLARY 2.4. *Suppose ν is an absolute norm. Then*

$$[\mu/\nu](E_{ij}) = \frac{\mu(e_i^{(m)})}{\nu(e_j^{(n)})}.$$

COROLLARY 2.5. *Let μ and ν be absolute norms on $\mathbf{F}^m$ and $\mathbf{F}^n$, respectively. Then*

$$|a_{ij}| \leq [\mu/\nu](A)[\mu/\nu](E_{ij})^{-1}$$

for all $A = (a_{ij}) \in \mathbf{F}^{mn}$, $1 \leq i \leq m$ and $1 \leq j \leq n$.

Proof. Let $A \in \mathbf{F}^{mn}$. Then, by Proposition 2.1 and Corollary 2.4,

$$[\mu/\nu](A) \geq \left| \frac{e_i^{(m)}}{\mu^D(e_i^{(m)})} A \frac{e_j^{(n)}}{\nu(e_j^{(n)})} \right| = |a_{ij}| \frac{\mu(e_i^{(m)})}{\nu(e_j^{(n)})} = |a_{ij}|[\mu/\nu](E_{ij})^{-1},$$

and so the corollary follows. ∎

3. Fractional Schur products of nonnegative matrices. In this section, we investigate the behavior of vector norms and operator norms on a *fractional Schur product* of *nonnegative* vectors and matrices, where a matrix $A \in \mathbf{F}^{mn}$ is called nonnegative if A is entrywise nonnegative.

For a nonnegative number t, the tth *Schur power* of a nonnegative matrix $A = (a_{ij})$ in $\mathbf{F}^{mn}$, denoted by $A^{[t]}$, is defined as $A^{[t]} = (a_{ij}^t)$.

Let $\mathcal{N}$ be a norm on $\mathbf{F}^{mn}$. A positive number C is called a *fractional Schur submultiplicativity factor* for the norm $\mathcal{N}$ if for any positive integer k and all positive numbers $\delta_1, \ldots, \delta_k$ such that $\sum_{i=1}^k \delta_i \geq 1$ we have

$$\mathcal{N}(A_1^{[\delta_1]} \circ \cdots \circ A_k^{[\delta_k]}) \leq C^{\delta - 1} \prod_{i=1}^k \mathcal{N}(A_i)^{\delta_i} \tag{4}$$

for all nonnegative matrices $A_1, \ldots, A_k$, where $\delta = \sum_{i=1}^k \delta_i$. If $C = 1$ in (4), then $\mathcal{N}$ is called *fractional Schur submultiplicative*. If (4) holds for fixed $k = 2$ and $\delta_1 = \delta_2 = 1$ and all nonnegative matrices A_1, A_2 in $\mathbf{F}^{mn}$, we call C a *nonnegative Schur submultiplicativity factor* for $\mathcal{N}$. In this case, if $C = 1$, then $\mathcal{N}$ is called *nonnegative Schur submultiplicative*.

Note that $(A, B) \mapsto \mathcal{N}(A \circ B)$ is a continuous function on $\mathbf{F}^{mn} \times \mathbf{F}^{mn}$ and $\Omega \times \Omega$ is a compact set in the usual Euclidean topology, where $\Omega = \{A \,:\, \mathcal{N}(A) = 1\}$. Thus one can define

$$S_{\mathcal{N}}^+ := \max\{\mathcal{N}(A \circ B) : \mathcal{N}(A) = \mathcal{N}(B) = 1,\ A \geq 0, B \geq 0\}. \tag{5}$$

It is not hard to show that $S_{\mathcal{N}}^+$ is the least nonnegative Schur submultiplicativity factor for $\mathcal{N}$ (see [G90] and [AG90], etc.).

The purpose of this section is to find the least fractional Schur submultiplicativity factor for an operator norm. In particular, a necessary and sufficient condition for $[\mu/\nu]$ to be nonnegative Schur submultiplicative is given (see Corollary 3.7). Further, we give a short proof for an inequality (see (14)) due to [KO85] and [EJS88] (see Proposition 3.8).

We shall identify $\mathbf{F}^{n1}$ with $\mathbf{F}^n$. If μ is a norm on $\mathbf{F}^n$ and ν is the absolute value on $\mathbf{F}$, then it is easily shown that $[\mu/\nu] = \mu$.

We begin with results on $\mathbf{F}^n$. The following lemma is essentially shown in the course of the proof of [JN91, Theorem 1]. For the sake of completeness, we give a short proof as follows.

LEMMA 3.1. *Let μ be a quasi-monotonic norm on $\mathbf{F}^n$. Then, for $x, y \geq 0$ in $\mathbf{F}^n$ and $0 \leq \delta \leq 1$,*

$$\mu(x^{[\delta]} \circ y^{[1-\delta]}) \leq \mu(x)^\delta \mu(y)^{1-\delta}.$$

Proof. First, we assume that $\mu(x) = \mu(y) = 1$ and all the coordinates of x and y are positive. By the well-known weighted mean-value inequality, we have

$$x_i^\delta y_i^{1-\delta} \leq \delta x_i + (1 - \delta) y_i$$

for $i = 1, \ldots, n$. It follows that

$$x^{[\delta]} \circ y^{[1-\delta]} \leq \delta x + (1 - \delta) y.$$

Since μ is quasi-monotonic, we have

$$\begin{aligned}\mu(x^{[\delta]} \circ y^{[1-\delta]}) &\le \mu(\delta x + (1-\delta)y) \\ &\le \delta\mu(x) + (1-\delta)\mu(y) = 1.\end{aligned}$$

So, for x, y in $\mathbf{C}^n$ with positive entries, we have

$$\mu\left(\left(\frac{x}{\mu(x)}\right)^{[\delta]} \circ \left(\frac{y}{\mu(y)}\right)^{[1-\delta]}\right) \le 1.$$

Thus the inequality holds for vectors with positive entries. By continuity, the lemma is true for $x \ge 0$ and $y \ge 0$ in $\mathbf{C}^n$. □

For a given norm $\mathcal{N}$ on $\mathbf{F}^{mn}$, we define

$$C_{\mathcal{N}} := \max\left\{\mathcal{N}(E_{ij})^{-1} : 1 \le i \le m, 1 \le j \le n\right\}. \tag{6}$$

In particular, for a norm μ on $\mathbf{F}^m$,

$$C_\mu = \max\left\{\mu(e_i^{(m)})^{-1} : 1 \le i \le m\right\}. \tag{7}$$

LEMMA 3.2. *Let μ be a quasi-monotonic norm on $\mathbf{F}^n$ and suppose $t \ge 1$.*
(a) *Let $C \ge C_\mu$. Then, for all nonnegative $x \in \mathbf{F}^n$, we have*

$$\mu(x^{[t]}) \le C^{t-1}\mu(x)^t. \tag{8}$$

(b) *If $t > 1$, then $C = C_\mu$ is the least positive number such that* (8) *holds.*
Proof.
(a) Obviously, the conclusion is true if $x = 0$.

For nonzero $x \in \mathbf{F}^n$, it is sufficient to prove (8) with $C = C_\mu$. First, assume that $\mu(x) = 1$. Since $0 \le x_i e_i \le x$, we have $x_i\mu(e_i) \le 1$, and so $x_i \le C_\mu$ for $i = 1, \dots, n$. This implies that

$$x^{[t]} \le C_\mu^{t-1}x.$$

Since μ is quasi-monotonic, it follows that

$$\mu(x^{[t]}) \le C_\mu^{t-1}\mu(x) = C_\mu^{t-1},$$

so the conclusion is true when $\mu(x) = 1$. Applying the proven result to $x/\mu(x)$ for $x \ne 0$, we have proved (a).

(b) Suppose that C is any positive number such that (8) holds. Let $x = e_i^{(n)}$ $(1 \le i \le n)$. By (8), we have

$$\mu(e_i^{(n)}) \le C^{t-1}\mu(e_i^{(n)})^t.$$

Since $t > 1$, the preceding inequality implies that

$$\mu(e_i^{(n)})^{-1} \le C$$

for $1 \le i \le n$. By (6), it follows that $C_\mu \le C$. □

PROPOSITION 3.3. *Let μ be a quasi-monotonic norm on $\mathbf{F}^n$. Suppose k is a positive integer and $\delta_1, \dots, \delta_k$ are nonnegative numbers such that $\delta = \sum_{i=1}^k \delta_i \ge 1$.*

(a) *Let $C \geq C_\mu$. Then, for any nonnegative $x_1, \ldots, x_k$ in $\mathbf{F}^n$, we have*

$$\mu(x_1^{[\delta_1]} \circ \cdots \circ x_k^{[\delta_k]}) \leq C^{\delta-1} \prod_{i=1}^{k} (\mu(x_i))^{\delta_i}. \tag{9}$$

(b) *If $\delta = \sum_{i=1}^{k} \delta_i > 1$, then $C = C_\mu$ is the least positive number such that* (9) *holds.*

Thus C_μ is the smallest fractional submultiplicativity factor for the norm μ.

Proof.

(a) It is sufficient to prove (9) with $C = C_\mu$. First, we prove the case of $\delta = 1$.

If $k = 1$, the conclusion is obvious. Suppose $k \geq 2$ and

$$y = (x_1^{[\delta_1]} \circ \cdots \circ x_{k-1}^{[\delta_{k-1}]})^{\frac{1}{1-\delta_k}}.$$

Since $0 \leq \delta_k \leq 1$, Lemma 3.1 ensures that

$$\mu(x_1^{[\delta_1]} \circ \cdots \circ x_k^{[\delta_k]}) = \mu(y^{[1-\delta_k]} \circ x_k^{[\delta_k]}) \leq \mu(y)^{1-\delta_k} \mu(x_k)^{\delta_k}. \tag{10}$$

By induction on k, we have

$$\mu(y) = \mu(x_1^{[\frac{\delta_1}{1-\delta_k}]} \circ \cdots \circ x_{k-1}^{[\frac{\delta_{k-1}}{1-\delta_k}]}) \leq \prod_{i=1}^{k-1} \mu(x_i)^{\frac{\delta_i}{1-\delta_k}}. \tag{11}$$

By (10) and (11), it follows that

$$\mu(x_1^{[\delta_1]} \circ \cdots \circ x_k^{[\delta_k]}) \leq \prod_{i=1}^{k} \mu(x_i)^{\delta_i}. \tag{12}$$

Now, assume that $\delta \geq 1$. We have

$$\sum_{i=1}^{k} \frac{\delta_i}{\delta} = 1.$$

By (12) and Lemma 3.2, it follows that

$$\begin{aligned} \mu(x_1^{[\delta_1]} \circ \cdots \circ x_k^{[\delta_k]}) &= \mu((x_1^{[\frac{\delta_1}{\delta}]} \circ \cdots \circ x_k^{[\frac{\delta_k}{\delta}]})^{[\delta]}) \\ &\leq C_\mu^{\delta-1} \mu(x_1^{[\frac{\delta_1}{\delta}]} \circ \cdots \circ x_k^{[\frac{\delta_k}{\delta}]})^\delta \leq C_\mu^{\delta-1} \prod_{i=1}^{k} (\mu(x_i)^{\frac{\delta_i}{\delta}})^\delta \\ &= C_\mu^{\delta-1} \prod_{i=1}^{k} \mu(x_i)^{\delta_i}. \end{aligned}$$

So, (a) is proved.

(b) Suppose that C is an arbitrary positive number such that (9) holds. Let each x_j $(1 \leq j \leq k)$ in (9) be $e_i^{(n)}$. By (9), we can obtain

$$\mu(e_i^{(n)}) \leq C^{\delta-1} \mu(e_i^{(n)})^\delta.$$

Since $\delta > 1$, we have

$$\mu(e_i^{(n)})^{-1} \le C$$

for $1 \le i \le n$. By (7), one has that $C_\mu \le C$. Therefore, C_μ is the least positive number such that (9) holds for the given $\delta_1, \ldots, \delta_k$. □

Next we study the fractional Schur products of nonnegative matrices on $\mathbf{F}^{mn}$. When the underlying norms μ and ν are semiabsolute and absolute, respectively, we prove that $C_{[\mu/\nu]}$ is the least fractional Schur submultiplicativity factor for $[\mu/\nu]$.

LEMMA 3.4. *Let μ and ν be norms on $\mathbf{F}^m$ and $\mathbf{F}^n$, respectively. If ν is an absolute norm, then*

$$C_{[\mu/\nu]} = \max\left\{ \frac{\nu(e_j^{(n)})}{\mu(e_i^{(m)})} : 1 \le i \le m, 1 \le j \le n \right\}.$$

Proof. It follows immediately by Corollary 2.4. □

Since $[\mu/\nu]$ is a quasi-monotonic norm on $\mathbf{F}^{mn}$ if μ is semiabsolute and ν is absolute (see Remark 2.3(b)), we can immediately obtain the following theorem by applying Proposition 3.3 to the norm $[\mu/\nu]$ on $\mathbf{F}^{mn}$.

THEOREM 3.5. *Let μ be a semiabsolute norm on $\mathbf{F}^m$ and let ν be an absolute norm on $\mathbf{F}^n$. Suppose k is a positive integer and $\delta_1, \ldots, \delta_k$ are nonnegative numbers such that $\delta = \sum_{i=1}^k \delta_i \ge 1$.*

(a) *Let $C \ge C_{[\mu/\nu]}$. Then, for any nonnegative $A_1, \ldots, A_k$ in $\mathbf{F}^{mn}$, we have*

$$[\mu/\nu](A_1^{[\delta_1]} \circ \cdots \circ A_k^{[\delta_k]}) \le C^{\delta-1} \prod_{i=1}^k [\mu/\nu](A_i)^{\delta_i}. \tag{13}$$

(b) *If $\delta = \sum_{i=1}^k \delta_i > 1$, then $C = C_{[\mu/\nu]}$ is the least positive number such that* (13) *holds.*

Thus $C_{[\mu/\nu]}$ is the smallest Schur fractional submultiplicativity factor for the norm $[\mu/\nu]$.

Remark 3.6.

(a) Note that we have proved more than that $C_{[\mu/\nu]}$ is the smallest fractional Schur submultiplicativity factor for the quasi-monotonic norm $[\mu/\nu]$ since, in Proposition 3.3 and Theorem 3.5, $\delta_1, \delta_2, \ldots, \delta_k$ are fixed nonnegative numbers. In particular, under the assumptions of Proposition 3.3 and Theorem 3.5 on norms, by putting $k = 2$ and $\delta_1 = \delta_2 = 1$, we see that the least nonnegative Schur submultiplicativity factor equals the least fractional Schur submultiplicativity factor.

(b) Suppose that μ is a weighted l_p norm on $\mathbf{F}^m$ with positive weight vector $x = (u_1, \ldots, u_m)$ and ν is a weighted l_q norm on $\mathbf{F}^n$ with positive weight vector $(v_1, \ldots, v_n)$, where $1 \le p, q \le \infty$. Then (6) ensures that

$$C_{[\mu/\nu]} = \max\left\{ \frac{v_j}{u_i} \ : \ 1 \le i \le m, 1 \le j \le n \right\}.$$

In particular, if μ is the l_p norm and ν is the l_q norm, then $[\mu/\nu]$ is nonnegative Schur submultiplicative and 1 is the least fractional Schur submultiplicativity factor for $[\mu/\nu]$.

By Theorem 3.5, (6), and Lemma 3.4 the following corollary follows immediately.

COROLLARY 3.7. *Let μ be a semiabsolute norm on $\mathbf{F}^m$ and ν be an absolute norm on $\mathbf{F}^n$. Then $C_{[\mu/\nu]}$ is a nonnegative Schur submultiplicativity factor.*

Further, the following statements are equivalent:

(a) $[\mu/\nu]$ *is nonnegative Schur submultiplicative*;
(b) $[\mu/\nu](E_{ij}) \geq 1$ *for* $1 \leq i \leq m, 1 \leq j \leq n$;
(c) $\mu(e_i^{(m)}) \geq \nu(e_j^{(n)})$ *for* $1 \leq i \leq m, 1 \leq j \leq n$.

In the following, we use $\rho(A)$ to denote the spectral radius of a square matrix A. The following proposition is due to [KO85] and [EJS88]. As an application of Theorem 3.5, we give a short proof of the result as follows.

PROPOSITION 3.8. *Suppose* $k \geq 1$. *Let* $A_1, \ldots, A_k \in \mathbf{F}^{nn}$ *be nonnegative. Then, for any positive numbers* $\delta_1, \ldots, \delta_k$ *with* $\sum_{i=1}^k \delta_i \geq 1$, *we have*

$$\rho(A_1^{[\delta_1]} \circ \cdots \circ A_k^{[\delta_k]}) \leq \prod_{i=1}^k \rho(A_i)^{\delta_i}. \tag{14}$$

Proof. Note that in [A] (for the irreducible nonnegative matrices) and in [HHSW, Theorem 2.2] it is proved that, for all nonnegative matrices $A \in \mathbf{F}^{nn}$,

$$\rho(A) = \inf_{X \in \mathcal{X}} \mathcal{N}(XAX^{-1}), \tag{15}$$

where $\mathcal{X}$ denotes the group of all $n \times n$ positive diagonal matrices and $\mathcal{N}$ is the operator norm on $\mathbf{F}^{nn}$ induced by some l_p norm on $\mathbf{F}^n$. By Remark 3.6 and Theorem 3.5, for any $X_1, \ldots, X_k$ in $\mathcal{X}$, we have

$$\begin{aligned} &\mathcal{N}((X_1^{\delta_1} \cdots X_k^{\delta_k})(A_1^{[\delta_1]} \circ \cdots \circ A_k^{[\delta_k]})(X_1^{\delta_1} \cdots X_k^{\delta_k})^{-1}) \\ &= \mathcal{N}((X_1 A_1 X_1^{-1})^{[\delta_1]} \circ \cdots \circ (X_k A_k X_k^{-1})^{[\delta_k]}) \\ &\leq \mathcal{N}(X_1 A_1 X_1^{-1})^{\delta_1} \cdots \mathcal{N}(X_k A_k X_k^{-1})^{\delta_k}. \end{aligned}$$

Let X_i $(i = 1, \ldots, k)$ run over the set $\mathcal{X}$. By (15), it follows that (14) holds. ∎

4. Schur submultiplicativity. Let $\mathcal{N}$ be a norm on $\mathbf{F}^{mn}$. Let C be a positive number. We call C a *Schur submultiplicativity factor* for the norm $\mathcal{N}$ if

$$\mathcal{N}(A \circ B) \leq C\mathcal{N}(A)\mathcal{N}(B) \tag{16}$$

holds for all $A, B \in \mathbf{F}^{mn}$. The submultiplicativity factor for a matrix norm for the usual matrix multiplication was investigated in [GS79], [GS82], [GS83a], and [GS83b], etc. In the terminology of [G90], [AG90], [AGL92], [AG93], [AGL93a], and [AGL93b], the number C would be called a multiplicativity factor for the algebra $\mathbf{F}^{mn}$ under Schur multiplication.

Similar to the definition of $S_{\mathcal{N}}^+$, one can define

$$S_{\mathcal{N}} := \max\{\mathcal{N}(A \circ B) : \mathcal{N}(A) = \mathcal{N}(B) = 1\}. \tag{17}$$

It is not hard to show that $S_{\mathcal{N}}$ is the least Schur submultiplicativity factor for $\mathcal{N}$. It is of interest to have some formulas for $S_{\mathcal{N}}$ so that it can be easily determined, say, computed in a finite number of steps. We are also interested in finding necessary and sufficient conditions for $S_{\mathcal{N}} \leq 1$. If this inequality holds, we call the norm $\mathcal{N}$ *Schur submultiplicative*.

In this section, we will obtain some simple expressions for $S_{[\mu/\nu]}$. We will also investigate the relationships among the numbers $S_{[\mu/\nu]}$, $S_{[\mu/\nu]}^+$, and $C_{[\mu/\nu]}$; see, e.g., Remark 4.5 and Theorem 4.7.

PROPOSITION 4.1. *Let μ and ν be norms on $\mathbf{F}^m$ and $\mathbf{F}^n$, respectively. Then*

$$C_{[\mu/\nu]} \leq \max\{[\mu/\nu](A \circ A) : [\mu/\nu](A) = 1\}$$

and

$$C_{[\mu/\nu]} \leq \max\{[\mu/\nu](A \circ \bar{A}) : [\mu/\nu](A) = 1\},$$

where $\bar{A}$ is the entrywise conjugate of A.

If μ and ν are conjugate invariant, then $\max\{[\mu/\nu](A \circ \bar{A}) : [\mu/\nu](A) = 1\} \leq S_{[\mu/\nu]}$.

Proof. Let $\eta = \max\{[\mu/\nu](A \circ A) : [\mu/\nu](A) = 1\}$. Since $[\mu/\nu](E_{ij}) \leq \eta[\mu/\nu](E_{ij})^2$, it follows that $1/[\mu/\nu](E_{ij}) \leq \eta$, where $1 \leq i \leq m, 1 \leq j \leq n$. By (6), the first inequality follows. Similarly, the second inequality is true.

If μ and ν are conjugate invariant, then $[\mu/\nu](\bar{A}) = [\mu/\nu](A)$ for all $A \in \mathbf{F}^{mn}$, so the last assertion follows. ∎

Note that if $\mathbf{F} = \mathbf{R}$ we have

$$\max\{[\mu/\nu](A \circ \bar{A}) : [\mu/\nu](A) = 1\} = \max\{[\mu/\nu](A \circ A) : [\mu/\nu](A) = 1\}.$$

Hence, for any norms μ on $\mathbf{R}^m$ and ν on $\mathbf{R}^n$, we have

$$C_{[\mu/\nu]} \leq \max\{[\mu/\nu](A^{[2]}) : A \in \mathbf{R}^{mn},\ [\mu/\nu](A) = 1\} \leq S_{[\mu/\nu]}.$$

In the rest of this section, we focus on the operator norms induced by semiabsolute or absolute norms.

THEOREM 4.2. *Let μ and ν be absolute norms on $\mathbf{F}^m$ and $\mathbf{F}^n$, respectively. Then*

$$S_{[\mu/\nu]} = \max\{[\mu/\nu](A \circ \bar{A}) : [\mu/\nu](A) = 1\}.$$

Proof. Since μ and ν are absolute, by Proposition 4.1 it follows that

$$\max\{[\mu/\nu](A \circ \bar{A}) : [\mu/\nu](A) = 1\} \leq S_{[\mu/\nu]}.$$

Conversely, denote

$$\eta = \max\{[\mu/\nu](A \circ \bar{A}) : [\mu/\nu](A) = 1\}.$$

Then, for any $A, B \in \mathbf{F}^{mn}$ with $[\mu/\nu](A) = [\mu/\nu](B) = 1$, by Proposition 2.2 and Theorem 3.5, we have

$$\begin{aligned}[\mu/\nu](A \circ B) &\leq [\mu/\nu](|A| \circ |B|) = [\mu/\nu]((|A|^{[2]})^{[1/2]} \circ (|B|^{[2]})^{[1/2]}) \\ &\leq [\mu/\nu](|A|^{[2]})^{1/2}\, [\mu/\nu](|B|^{[2]})^{1/2} = [\mu/\nu](A \circ \bar{A})^{1/2}\, [\mu/\nu](B \circ \bar{B})^{1/2} \\ &\leq \eta^{1/2} \cdot \eta^{1/2} = \eta.\end{aligned}$$

This shows that

$$S_{[\mu/\nu]} \leq \max\{[\mu/\nu](A \circ \bar{A}) : [\mu/\nu](A) = 1\}. \qquad ∎$$

Theorem 4.2 shows that the definition of $S_{[\mu/\nu]}$ can be simplified if the operator norm is induced by absolute norms. The following corollary follows immediately from Theorem 4.2.

COROLLARY 4.3. *Let μ and ν be absolute norms on $\mathbf{F}^m$ and $\mathbf{F}^n$, respectively. Then $[\mu/\nu]$ is Schur submultiplicative if and only if $[\mu/\nu](A \circ \bar{A}) \le [\mu/\nu](A)^2$ for all $A \in \mathbf{F}^{mn}$.*

In the remainder of this section, we will investigate the relationship between $S_{[\mu/\nu]}$ and $S^+_{[\mu/\nu]}$.

THEOREM 4.4. *Let μ and ν be absolute norms on $\mathbf{F}^m$ and $\mathbf{F}^n$, respectively. Then*

$$S_{[\mu/\nu]} \le \min\{m, n\} C_{[\mu/\nu]}.$$

Proof. Assume $A \in \mathbf{F}^{mn}$ with $[\mu/\nu](A) = 1$. Let $1 \le i \le n$ and let A_i be the matrix in $\mathbf{F}^{mn}$ such that the ith column of A_i is the same as the ith column of A and the other columns of A_i equal zero. For any $x \in \mathbf{F}^n$ with $\nu(x) \le 1$, we have $A_i x = A(x_i e_i^{(n)})$. Since ν is also monotonic, it follows that $\nu(x_i e_i^{(n)}) \le 1$, and so $[\mu/\nu](A_i) \le [\mu/\nu](A) = 1$. On the other hand, since μ is absolute, we have $[\mu/\nu](|A_i|) = [\mu/\nu](A_i)$. Hence, we have

$$\begin{aligned}
[\mu/\nu](A \circ \bar{A}) &= [\mu/\nu]\left(\sum_{i=1}^n |A_i|^{[2]}\right) \le \sum_{i=1}^n [\mu/\nu](|A_i|^{[2]}) \\
&\le C_{[\mu/\nu]} \sum_{i=1}^n [\mu/\nu](|A_i|)^2 = C_{[\mu/\nu]} \sum_{i=1}^n [\mu/\nu](A_i)^2 \\
&\le n C_{[\mu/\nu]},
\end{aligned}$$

where the second inequality follows from Theorem 3.5. By applying similar arguments to the rows of A, we can prove that $[\mu/\nu](A \circ \bar{A}) \le m C_{[\mu/\nu]}$. By Theorem 4.2 and the preceding arguments, the theorem follows. □

Remark 4.5. By (5), (17), Remark 3.6(a), and Theorem 4.4 we have the following relationships among $S_{[\mu/\nu]}$, $S^+_{[\mu/\nu]}$, and $C_{[\mu/\nu]}$ if μ and ν are absolute norms:

$$C_{[\mu/\nu]} = S^+_{[\mu/\nu]} \le S_{[\mu/\nu]} \le \min\{m, n\} C_{[\mu/\nu]}. \tag{18}$$

Thus, for the case $n = 1$, we have $S^+_\mu = C_\mu = S_\mu$, which is easily proved independently.

The following example of Bit-Shun Tam shows that $S^+_{[\mu/\nu]} = S_{[\mu/\nu]}$ need not be true when μ and ν are absolute norms.

Example 4.6. On $\mathbf{F}^2$, define $\mu(x) = \max\{|x_1|, |x_2|, 2(|x_1| + |x_2|)/3\}$. Let $\nu = \mu^D$. Then the unit norm ball of $\nu = \mu^D$ has

$$\mathcal{E} = \{\lambda e_i^{(2)} : i = 1, 2, |\lambda| = 1\} \cup \{(u_1, u_2)^t : u_1, u_2 \in \mathbf{F},\ |u_1| = |u_2| = 2/3\}$$

as its set of extreme points. By Proposition 2.1,

$$[\mu/\nu](B) = \max\{|x^* B y| : x, y \in \mathcal{E}\}.$$

Let $A = \begin{bmatrix} 1 & -1/2 \\ 1/2 & 1 \end{bmatrix}$. Then $[\mu/\nu](A) = 1$ and $[\mu/\nu](A \circ \bar{A}) \ge 10/9$, and so $S_{[\mu/\nu]} \ge 10/9$. But by Lemma 3.4 and Remark 3.6(a), we have $S^+_{[\mu/\nu]} = C_{[\mu/\nu]} = 1$.

Some sufficient conditions on μ and ν for $S_{[\mu/\nu]} = S^+_{[\mu/\nu]}$ are given in the following theorem.

THEOREM 4.7. *Let μ and ν be absolute norms on $\mathbf{F}^m$ and $\mathbf{F}^n$, respectively. If either μ or ν^D is a weighted l_∞ norm, then $S_{[\mu/\nu]} = S^+_{[\mu/\nu]} = C_{[\mu/\nu]}$.*

Proof. Since $S^+_{[\mu/\nu]} \le S_{[\mu/\nu]}$, by Remark 4.5, it is sufficient to prove that $S_{[\mu/\nu]} \le C_{[\mu/\nu]}$.

Assume that $\mu(x) = l_\infty(w \circ x)$ for all $x \in \mathbf{F}^m$, where $w \in \mathbf{F}^m$ is entrywise positive. Then $\mu^D(x) = l_1(w^{[-1]} \circ x)$ for all $x \in \mathbf{F}^m$. By Proposition 2.1, for all $A \in \mathbf{F}^{mn}$ we have

$$[\mu/\nu](A) = \max\{w_i|(e_i^{(m)})^t A y| : 1 \le i \le m, \nu(y) = 1\}.$$

Now, suppose $A \in \mathbf{F}^{mn}$ with $[\mu/\nu](A) = 1$ and $y \in \mathbf{F}^n$ with $\nu(y) = 1$. By Corollary 2.5, $|a_{ij}| \le C_{[\mu/\nu]}$ for $1 \le i \le m, 1 \le j \le n$. Then, for each i,

$$\begin{aligned} w_i|(e_i^{(m)})^t (A \circ \bar{A}) y| &\le w_i \sum_{j=1}^n |a_{ij}|^2 |y_j| \le C_{[\mu/\nu]} w_i \sum_{j=1}^n |a_{ij}||y_j| \\ &= C_{[\mu/\nu]} w_i |(e_i^{(m)})^t A (D_i y)| \le C_{[\mu/\nu]}, \end{aligned}$$

where D_i is a unitary diagonal matrix in $\mathbf{F}^{nn}$, and so $\nu(D_i y) = 1$ since ν is an absolute norm. So, $[\mu/\nu](A \circ \bar{A}) \le C_{[\mu/\nu]}$. By Theorem 4.2, it follows that $S_{[\mu/\nu]} \le C_{[\mu/\nu]}$.

If ν^D is a weighted l_∞ norm, the proof is similar. ∎

By Theorem 4.7, (6), and Lemma 3.4 the following corollary follows immediately.

COROLLARY 4.8. *Let μ and ν be absolute norms on $\mathbf{F}^m$ and $\mathbf{F}^n$, respectively. Suppose either μ or ν^D is a weighted l_∞ norm. Then the following are equivalent:*

(a) *$[\mu/\nu]$ is Schur submultiplicative*;

(b) *$[\mu/\nu](E_{ij}) \ge 1$ for $1 \le i \le m, 1 \le j \le n$;*

(c) *$\mu(e_i^{(m)}) \ge \nu(e_j^{(n)})$ for $1 \le i \le m, 1 \le j \le n$.*

Under the conditions of Corollary 4.8, (b) and (c) are equivalent (see Lemma 3.4). In general, neither (b) nor (c) implies (a) (see Example 4.6), even if μ and ν are both absolute and *permutation invariant* in the sense that $\mu(Px) = \mu(x)$ for all x and permutation matrices P. If the operator norm $[\mu/\nu]$ is also absolute, then obviously $S_{[\mu/\nu]} = S^+_{[\mu/\nu]}$. But the operator norm induced by absolute norm(s) need not be absolute, e.g., the operator norm induced by the l_2 norm. On the other hand, the inequality in (16) with $C = 1$ shows that $S_{[\mu/\nu]} = S^+_{[\mu/\nu]} = C_{[\mu/\nu]}$ if the operator norm is induced by the l_p and l_q norms, and further that they are all equal to 1 (see [Be77], [N89], [S11], [On84], [Ok87], and [HJ91, 5.5.15]). So, the absolute property of $[\mu/\nu]$ is not necessary for $S_{[\mu/\nu]} = S^+_{[\mu/\nu]}$.

The following questions arise naturally.

Questions 4.9.

(a) What is a necessary and sufficient condition on μ and ν for $S_{[\mu/\nu]} = S^+_{[\mu/\nu]}$?

(b) How does one determine $S_{[\mu/\nu]}$?

5. Limit of norms and Schur operator norms. Suppose μ is a norm on $\mathbf{F}^n$ and K is a nonempty closed cone (convex) in $\mathbf{F}^n$ with nonempty interior. The *Schur K-operator norm induced by μ* on $\mathbf{F}^n$, denoted by μ^K, is

$$\mu^K(x) = \max\{\mu(x \circ y) : y \in K, \mu(y) = 1\}, \quad x \in \mathbf{F}^n. \tag{19}$$

Two typical choices of K are $\mathbf{F}^n$ and the cone of all nonnegative vectors in $\mathbf{F}^n$. If $K = \mathbf{F}^n$, the induced norm μ^K is called a *Schur operator norm* on $\mathbf{F}^n$.

THEOREM 5.1. *Let μ be a norm on $\mathbf{F}^n$. Then $\lim_{t\to\infty} \mu(x^{[t]})^{1/t}$ exists and is equal to $l_\infty(x)$ for all $x \in \mathbf{F}^n$.*

Proof. By [HJ85], there exist finite positive constants C_1 and C_2 such that

$$C_1 l_\infty \le \mu \le C_2 l_\infty.$$

Obviously, $l_\infty(x^{[t]}) = l_\infty(x)^t$ for all $x \in \mathbf{F}^n$ and all positive t. Hence, for all $x \in \mathbf{F}^n$ and all positive t, it follows that

$$C_1^{1/t} l_\infty(x) \le \mu(x^{[t]})^{1/t} \le C_2^{1/t} l_\infty(x).$$

Let $t \to \infty$. The above inequalities show that $\lim_{t\to\infty} \mu(x^{[t]})^{1/t}$ exists and is equal to $l_\infty(x)$. ∎

The following theorem shows that $\mu^K(x)$ can be computed in a finite number of steps if K is the cone of all nonnegative vectors in $\mathbf{F}^n$, x is in K, and μ is quasi-monotonic.

THEOREM 5.2. *Let μ be a quasi-monotonic norm on $\mathbf{F}^n$. Let K be the cone of all nonnegative vectors in $\mathbf{F}^n$. Then, for all nonnegative $x \in \mathbf{F}^n$,*

$$\mu^K(x) = l_\infty(x).$$

Proof. If $x = 0$, the conclusion is trivial.

Assume that $x \neq 0$. For any $y \in K$ with $\mu(y) = 1$, we have

$$\mu(x \circ y) \le l_\infty(x)\mu(y) = l_\infty(x).$$

It follows that $\mu^K(x) \le l_\infty(x)$.

Suppose t is a positive integer. Then, by (19),

$$\frac{\mu(x^{[t]})}{\mu(x^{[t-1]})} = \frac{\mu(x \circ x^{[t-1]})}{\mu(x^{[t-1]})} \le \mu^K(x),$$

so

$$\mu(x^{[t]}) \le \mu^K(x)\mu(x^{[t-1]}), \quad t = 1, 2, \ldots.$$

It follows that

$$\mu(x^{[t]}) \le \mu^K(x)\mu(x^{[t-1]}) \le (\mu^K(x))^2 \mu(x^{[t-2]}) \le \cdots \le (\mu^K(x))^{t-1}\mu(x).$$

Therefore,

$$\mu(x^{[t]})^{1/t} \le (\mu^K(x))^{(t-1)/t}\mu(x)^{1/t}.$$

Let $t \to \infty$. By Theorem 5.1, we have $l_\infty(x) \le \mu^K(x)$. ∎

COROLLARY 5.3. *Let μ be a semiabsolute norm on $\mathbf{F}^m$ and ν be an absolute norm on $\mathbf{F}^n$. Let K be the cone of all nonnegative matrices in $\mathbf{F}^{mn}$. Then for all nonnegative $A \in \mathbf{F}^{mn}$ we have $[\mu/\nu]^K(A) = l_\infty(A)$.*

Proof. Since $[\mu/\nu]$ on $\mathbf{F}^{mn}$ is quasi-monotonic (Remark 2.3(b)), the assertion follows immediately by applying Theorem 5.2 to the space $\mathbf{F}^{mn}$. ∎

REFERENCES

[A] J. ALBRECHT, *Minimal norms of non-negative, irreducible matrices*, Linear Algebra Appl., 249 (1996), pp. 255–258.

[AG90] R. Arens and M. Goldberg, *Multiplicativity factors for seminorms*, J. Math. Anal. Appl., 146 (1990), pp. 469–481.

[AG93] R. Arens and M. Goldberg, *Quadrative seminorms and Jordan structures on algebras*, Linear Algebra Appl., 181 (1993), pp. 269–278.

[AG94] R. Arens and M. Goldberg, *Weighted l_∞ norms for matrices*, Linear Algebra Appl., 201 (1994), pp. 155–163.

[AGL92] R. Arens, M. Goldberg, and W. A. J. Luxemburg, *Multiplicativity factors for seminorms* II, J. Math. Anal. Appl., 170 (1992), pp. 401–413.

[AGL93a] R. Arens, M. Goldberg, and W. A. J. Luxemburg, *Multiplicativity factors for function norms*, J. Math. Anal. Appl., 177 (1993), pp. 368–385.

[AGL93b] R. Arens, M. Goldberg, and W. A. J. Luxemburg, *Multiplicativity factors for Orlicz space function norms*, J. Math. Anal. Appl., 177 (1993), pp. 386–411.

[AHJ87] T. Ando, R. A. Horn, and C. R. Johnson, *The singular values of a Hadamard product: A basic inequality*, Linear and Multilinear Algebra, 21 (1987), pp. 345–365.

[BSW61] F. L. Bauer, J. Stoer, and C. Witzgall, *Absolute and monotonic norms*, Numer. Math., 3 (1961), pp. 257–264.

[Ba63] F. L. Bauer, *Optimally scaled matrices*, Numer. Math., 5 (1963), pp. 73–87.

[Be77] G. Bennett, *Schur multipliers*, Duke Math. J., 44 (1977), pp. 603–639.

[EJS88] L. Elsner, C. R. Johnson, and J. A. Dias da Silva, *The Perron root of a weighted geometric mean of nonnegative matrices*, Linear and Multilinear Algebra, 24 (1988), pp. 1–13.

[FH77] C. H. FitzGerald and R. A. Horn, *On fractional Hadamard powers of positive definite matrices*, J. Math. Anal. Appl., 61 (1977), pp. 633–642.

[G89] M. Goldberg, *A note on monotonic and semi-monotonic matrix functions*, Linear and Multilinear Algebra, 24 (1989), pp. 223–226.

[G90] M. Goldberg, *Quasimonotonic functions on $\mathbf{C}^n$ and the mapping $f \mapsto f^+$*, Linear and Multilinear Algebra, 27 (1990), pp. 63–71.

[GS79] M. Goldberg and E. G. Straus, *Norm properties of C-numerical radii*, Linear Algebra Appl., 24 (1979), pp. 113–131.

[GS82] M. Goldberg and E. G. Straus, *Operator norms, multiplicativity factors, and C-numerical radii*, Linear Algebra Appl., 43 (1982), pp. 137–159.

[GS83a] M. Goldberg and E. G. Straus, *Multiplicativity of l_p norms for matrices*, Linear Algebra Appl., 52/53 (1983), pp. 351–360.

[GS83b] M. Goldberg and E. G. Straus, *Multiplicativity factors for C-numerical radii*, Linear Algebra Appl., 54 (1983), pp. 1–16.

[H90] R. A. Horn, *The Hadamard product*, in Proc. Symposia in Applied Mathematics, 40 (1990), pp. 87–169.

[HHSW] D. Hershkowitz, W. Huang, H. Schneider, and H. Weinberger, *Approximability by weighted norms of the structured and volumetric singular values of a class of nonnegative matrices*, SIAM J. Matrix Anal. Appl., 18 (1997), to appear.

[HJ85] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.

[HJ87] R. A. Horn and C. R. Johnson, *Hadamard and conventional submultiplicativity for unitarily invariant norms on matrices*, Linear and Multilinear Algebra, 21 (1987), pp. 91–106.

[HJ91] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[HM90] R. A. Horn and R. Mathias, *An analog of the Cauchy-Schwarz inequality for Hadamard products and unitarily invariant norms*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 481–498.

[Ho1889] L. O. Hölder, *Über einen Mittelwerthssatz*, Nachr. Königl. Ges. Wiss. und G.A. Univ. Göttingen, (1889), pp. 38–47.

[JN91] C. R. Johnson and P. Nylen, *Monotonicity properties of norms*, Linear Algebra Appl., 148 (1991), pp. 43–58.

[KO85] S. Karlin and F. Ost, *Some monotonicity properties of Schur powers of matrices and related inequalities*, Linear Algebra Appl., 68 (1985), pp. 47–65.

[MKS84] M. Marcus, K. Kidman, and M. Sandy, *Unitarily invariant generalized matrix norms and Hadamard products*, Linear and Multilinear Algebra, 16 (1984), pp. 197–213.

[MP96] B. Mond and J. E. Pečarič, *On an inequality for spectral radius*, Linear and Multilinear Algebra, 40 (1996), pp. 203–206.

[N89] P. Nylen, *Submultiplicativity of Matrix Products,* Ph.D. thesis, Clemson University, Clemson, SC, 1989.

[Ok87] K. Okubo, *Hölder-type norm inequalities for Schur products of matrices*, Linear Algebra Appl., 91 (1987), pp. 13–28.

[On84] S.-C. Ong, *On the Schur multiplier norm of matrices*, Linear Algebra Appl., 56 (1984), pp. 45–55.

[S11] J. Schur, *Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen*, J. Reine Angew. Math., 140 (1911), pp. 1–28.

[SS75] B. D. Saunders and H. Schneider, *Norms and Numerical Ranges in Finite Dimensions,* unpublished notes, 1975.

[W86] M. E. Walter, *On the norm of a Schur product*, Linear Algebra Appl., 79 (1986), pp. 209–213.

[Zh88] F.-Z. Zhang, *Another proof of a singular value inequality concerning Hadamard products of matrices*, Linear and Multilinear Algebra, 22 (1988), pp. 307–311.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, pp. 348–359, April 1997

007

MATRIX ANALYSIS OF A TWO-STAGE-SPLITTING ITERATION FOR MAXIMUM PENALIZED LIKELIHOOD ESTIMATION*

S. YU†, G. A. LATHAM†, AND R. S. ANDERSSEN†

Abstract. Motivated by operator splitting, a new two-stage-splitting iteration is proposed for the solution of maximum penalized likelihood estimation (MPLE) problems. The resulting algorithm, called two-step-late (TSL), is as practical and as easily implemented as the one-step-late (OSL) algorithm. Matrix analysis is applied to compare the rates of convergence of the TSL and OSL algorithms. It is proved that under quite general conditions for which OSL and TSL converge to the same solution the rate of convergence of TSL exceeds that of two steps of OSL, which is its computational counterpart. Numerical experimentation can then be used to check the range of the smoothing parameter for which these proofs hold.

Key words. EM, OSL, perturbation and spectral theory, two-stage-splitting

AMS subject classifications. 15A18, 65U05, 92C55

PII. S0895479896260160

1. Introduction.

The expectation maximization (EM) methodology [4] is a general approach for maximizing a likelihood or posterior distribution rather than a specific algorithm. The common strand in problems where this approach is applicable is the notion of "incomplete data," which includes the conventional sense of "missing data." The EM methodology demonstrates its strength in such situations [14]. The resulting algorithms are extremely simple and remain practical for large problems where other approaches do not appear to be feasible.

Along with the recent growth of interest in the use of Bayesian and penalized likelihood methods, the EM methodology has been increasingly used to implement these methods. Green [7] has discussed the EM methodology for MPLE problems. He proposed an OSL algorithm which is often much easier to implement than the original EM algorithm and converges slightly faster than it does. The OSL algorithm can also be applied to posterior probability problems [6].

This paper proposes a new two-stage-splitting iteration. The motivation is the iterative methods based on operator splitting [5]. They are popular in other contexts such as the solution of the Navier–Stokes equations in fluid dynamics [5] since, through a judicious splitting of the underlying operator, iterative schemes can be constructed which have improved rates of convergence as well as other desirable numerical properties. In fact, the splitting introduces flexibility into the way iterative methods can be defined. For the new algorithm, called two-step-late (TSL), the nonlinear operator, whose kernel maximizes the MPLE functional, is split into two parts as outlined below. It is then desirable to construct an appropriate mathematical framework to evaluate the numerical performance of the TSL algorithm, such as a comparison of its rate of convergence with its computational counterpart—two steps of OSL. Matrix perturbation and spectral theories are invoked for this purpose. It is proved that under quite realistic conditions if the OSL and TSL methods converge they converge to the same point, and the rate of convergence of TSL exceeds that of two steps of OSL. A general convergence theory for the TSL algorithm, as well as that of the

*Received by the editors January 4, 1996; accepted for publication (in revised form) by G. Cybenko May 8, 1996.

http://www.siam.org/journals/simax/18-2/26016.html

†Centre for Mathematics and its Applications, Australian National University, Canberra ACT 0200, Australia (shihong@maths.anu.edu.au, gal851@cscgpo.anu.edu.au, boba@cbr.dms.csiro.au).

OSL algorithm, remains an open question, but for MPLE problems convergence will depend heavily on the penalty employed. However, numerical experiments indicate that for such problems with a small penalty the OSL and TSL algorithms converge.

Remark 1. As shown by Lange [12], in any implementation of OSL (and hence by default TSL), convergence is greatly improved through the judicious introduction of a line search strategy into the maximization. Within any multistage iteration such as TSL one has greater flexibility for the introduction of line searches. Numerical experimentation, which is reported elsewhere [17], shows that the numerical performance of TSL with line search compares favorably with OSL with line search.

An explanation of the EM methodology and its use for the solution of MPLE problems is given in section 2. Two-stage-splitting is introduced in section 3 along with the definition of the TSL algorithm. Three theorems (Theorems 1, 2, and 3) with three different sets of conditions are proved to demonstrate the superiority of the rate convergence of TSL over that of two steps of OSL. Theorem 1 is based on perturbation results for eigenvalues and eigenvectors. Theorem 2 is derived using estimates of the spectral norm; Theorem 3 is proved using matrix commutativity and, in particular, applies to the one-dimensional case. In section 4, a numerical experiment is provided to illustrate the theoretical results in section 3. The example used is that of a multinomial sample as considered by other authors [4, 7].

2. The OSL algorithm and MPLE problems. In many applications, the given data (measurements) $\boldsymbol{y}$ can be viewed (interpreted) as indirect (incomplete) measurements of the underlying phenomenon of interest $\boldsymbol{x}$. The goal therefore becomes one of recovering information about the phenomenon $\boldsymbol{x}$ from the available data $\boldsymbol{y}$. There are various mathematical and statistical ways in which this can be done [1, 8]. Here, attention focuses on the maximum penalized likelihood formalism for incomplete data and its solution via the OSL algorithm.

Let $f(\boldsymbol{x}|\boldsymbol{\theta})$ denote the sampling density of the complete data $\boldsymbol{x}$ which depends on some parameters $\boldsymbol{\theta}$ and $g(\boldsymbol{y}|\boldsymbol{\theta})$ the corresponding sample density of the incomplete data $\boldsymbol{y}$. The aim of the ML (maximum likelihood) methodology is to determine the $\boldsymbol{\theta}$ which maximizes $\log g(\boldsymbol{y}|\boldsymbol{\theta})$ for a given $\boldsymbol{y}$ but by making essential use of the family $f(\boldsymbol{x}|\boldsymbol{\theta})$. The key to the derivation of the EM algorithm is the decision as to how to make use of the underlying family $f(\boldsymbol{x}|\boldsymbol{\theta})$.

The underlying heuristic, proposed in [4], is the observation that one would like to determine $\boldsymbol{\theta}$ so as to maximize $\log f(\boldsymbol{x}|\boldsymbol{\theta})$, except that $\boldsymbol{x}$ is unknown. Therefore it is necessary to work with some appropriate approximation of $\log f(\boldsymbol{x}|\boldsymbol{\theta})$, such as

$$Q(\boldsymbol{\theta}'|\boldsymbol{\theta}) = E(\log f(\boldsymbol{x}|\boldsymbol{\theta}')|\boldsymbol{y}, \boldsymbol{\theta}), \qquad \boldsymbol{\theta}' = \boldsymbol{\theta}.$$

In practice, one determines the $\boldsymbol{\theta}^*$ which maximizes $Q(\boldsymbol{\theta}|\boldsymbol{\theta})$ by iteratively maximizing $Q(\bar{\boldsymbol{\theta}}|\boldsymbol{\theta}_0)$ as a function of $\bar{\boldsymbol{\theta}}$ with respect to the current estimate $\boldsymbol{\theta}_0$ of $\boldsymbol{\theta}^*$. There is clearly some flexibility in how this might be done depending on how one decides to formulate the iterative process. The approach of [4] is to specify the iteration so as to retain the essential structure of the EM algorithm for exponential families. In this way, one obtains the iterative process $\boldsymbol{\theta}^{(n)} \to \boldsymbol{\theta}^{(n+1)}$ for $n \geq 0$ defined by an *E-step*: for the current $\boldsymbol{\theta}^{(n)}$, *estimate*

$$Q(\cdot|\boldsymbol{\theta}^{(n)}) = E(\log f(\boldsymbol{x}|\cdot)|\boldsymbol{y}, \boldsymbol{\theta}^{(n)}),$$

where the notation implies that the conditional expectation, given $\boldsymbol{y}$ and $\boldsymbol{\theta}^{(n)}$, is applied to $\log f(\boldsymbol{x}|\cdot)$ as if it were only a function of $\boldsymbol{x}$ (i.e., with the $\boldsymbol{\theta}$ held constant).

This is followed by an *M-step*: determine the $\boldsymbol{\theta}^{(n+1)}$ which *maximizes* $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(n)})$. Formally, these steps are implemented by replacing an initial $\boldsymbol{\theta}^{(0)}$ by that $\boldsymbol{\theta}^{(1)}$ which maximizes $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)})$. This replacement is continued until convergence is achieved, though practical considerations require that this process be stopped at an appropriate earlier stage.

The justification for this iterative process is that the $\boldsymbol{\theta}^*$ which the above EM method aims to approximate in fact does maximize $L(\boldsymbol{\theta})$ (defined below). In addition, the iterates monotonically increase the incomplete data likelihood and thereby converge [2, 3, 15].

If $k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ denotes the conditional density of $\boldsymbol{x}$ given $\boldsymbol{y}$, then $Q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ takes the form [4]

$$Q(\boldsymbol{\theta}'|\boldsymbol{\theta}) = L(\boldsymbol{\theta}') + H(\boldsymbol{\theta}'|\boldsymbol{\theta}), \tag{1}$$

where

$$L(\boldsymbol{\theta}') = \log g(\boldsymbol{y}|\boldsymbol{\theta}') \quad \text{and} \quad H(\boldsymbol{\theta}'|\boldsymbol{\theta}) = E(\log k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}')|\boldsymbol{y},\boldsymbol{\theta}).$$

For MPLE, instead of choosing $Q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ as the approximation to $\log f(\boldsymbol{x}|\boldsymbol{\theta})$, one takes the penalized objective function

$$Q(\boldsymbol{\theta}'|\boldsymbol{\theta}) - \lambda J(\boldsymbol{\theta}'), \tag{2}$$

where $J(\boldsymbol{\theta})$ is the penalty and $\lambda > 0$ is the smoothing (regularization) parameter. In this way, the EM method for MPLE corresponds to replacing $Q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ by (2) in the above definition of EM. In addition, the corresponding M-step reduces iteratively to solving, with respect to a given starting solution $\boldsymbol{\theta}^{(0)}$, the (nonlinear) equation

$$D_{10}Q(\boldsymbol{\theta}^{(n+1)}|\boldsymbol{\theta}^{(n)}) - \lambda DJ(\boldsymbol{\theta}^{(n+1)}) = 0, \tag{3}$$

where D denotes the derivative operator and, more generally, $D_{ij}F(\boldsymbol{\zeta}|\boldsymbol{\eta})$ denotes $D^i_{\boldsymbol{\zeta}}D^j_{\boldsymbol{\eta}}F(\boldsymbol{\zeta}|\boldsymbol{\eta})$.

For $\lambda > 0$ and a general $J(\boldsymbol{\theta})$, (3) could involve considerable computational effort as encountered in [6]. Hence, the OSL algorithm is defined as a modification of (3); namely,

$$D_{10}Q(\boldsymbol{\theta}^{n+1}|\boldsymbol{\theta}^n) - \lambda DJ(\boldsymbol{\theta}^n) = 0, \qquad n = 0, 1, \ldots. \tag{4}$$

Although in many applications of the EM methodology physical constraints imposed on the necessary optimization require consideration of the more general Kuhn–Tucker conditions, the scope of this paper is restricted to the simpler case in which all the iterates (for $\boldsymbol{\theta}$) lie in the interior of the parameter space and, therefore, equations such as (3) and (4) hold.

Let M denote the OSL map $M : \boldsymbol{\theta}^n \mapsto \boldsymbol{\theta}^{n+1}$; i.e., $\boldsymbol{\theta}^{n+1} = M(\boldsymbol{\theta}^n)$, where $\boldsymbol{\theta}^{n+1}$ is defined implicitly by (4). Differentiating (4) with respect to $\boldsymbol{\theta}^n$ yields

$$D_{11}Q(\boldsymbol{\theta}^{n+1}|\boldsymbol{\theta}^n) + D_{20}Q(\boldsymbol{\theta}^{n+1}|\boldsymbol{\theta}^n)DM(\boldsymbol{\theta}^n) - \lambda D_2 J(\boldsymbol{\theta}^n) = 0.$$

If the sequence $\{\boldsymbol{\theta}^n\}_{n\geq 0}$ converges to $\hat{\boldsymbol{\theta}}$, then by continuity

$$D_{11}Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) + D_{20}Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})DM(\hat{\boldsymbol{\theta}}) - \lambda D_2 J(\hat{\boldsymbol{\theta}}) = 0,$$

which can be rewritten as $DM(\hat{\boldsymbol{\theta}}) = (B + C)^{-1}(C - \lambda K)$, where [7, p. 445]

$$C = D_{11}Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) = D_{11}H(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) = -D_{20}H(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}),$$
$$B = -D_2 L(\hat{\boldsymbol{\theta}}), \quad \text{and} \quad K = D_2 J(\hat{\boldsymbol{\theta}}).$$

Note that B, C, and K depend implicitly on λ through their dependence on $\hat{\boldsymbol{\theta}}(\lambda)$.

In order to illustrate how the EM and OSL algorithms work on real problems, we consider a concrete example; namely, the multinomial sample problem considered in [4] and [7]. This example illustrates the essential nature of EM: the data set is completed using the current estimate of the parameter, and then using this completed data a new (MPL) estimate of the parameter is generated (cf. (7) below). This process is then iterated. Additional concrete illustrations can be found in section 8.8 of [10].

Example. Let the observed data consist of

$$\boldsymbol{y} = (y_1, y_2, y_3, y_4)^T = (125, 18, 20, 34)^T,$$

which are drawn from the multinomial distribution

$$M\left(197; \frac{1}{2} + \frac{1}{4}\theta, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{1}{4}\theta\right)$$

for some $\boldsymbol{\theta}$ with $0 \le \boldsymbol{\theta} \le 1$. The hypothesized complete data, which are drawn from

$$M\left(197; \frac{1}{2}, \frac{1}{4}\theta, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{1}{4}\theta\right),$$

consist of $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5)^T$, where

$$y_1 = x_1 + x_2, \quad y_2 = x_3, \quad y_3 = x_4, \quad \text{and} \quad y_4 = x_5.$$

For this example, one easily computes that

$$\hat{x}_1(\boldsymbol{y}, \theta) = 2y_1/(2+\theta), \quad \hat{x}_2(\boldsymbol{y}, \theta) = y_1\theta/(2+\theta),$$

$$\hat{x}_3(\boldsymbol{y}, \theta) = y_2, \quad \hat{x}_4(\boldsymbol{y}, \theta) = y_3, \quad \hat{x}_5(\boldsymbol{y}, \theta) = y_4,$$

$$\log f(\boldsymbol{x}|\theta') = x_1 \log \frac{1}{2} + (x_2 + x_5) \log \frac{\theta'}{4} + (x_3 + x_4) \log \frac{1-\theta'}{4} + \log \frac{197!}{\Pi_{j=1}^5 x_j!},$$

$$L(\theta') = y_1 \log \frac{2+\theta'}{4} + (y_2 + y_3) \log \frac{1-\theta'}{4} + y_4 \log \frac{\theta'}{4} + \log \frac{197!}{\Pi_{j=1}^4 y_j!},$$

where $\hat{x}_j, j = 1, 2, \ldots, 5$, are estimates (expectations) of x_j given $\boldsymbol{y}$ and θ.

It follows that

$$Q(\theta'|\theta) = \hat{x}_1 \log \frac{1}{2} + (\hat{x}_2 + \hat{x}_5) \log \frac{\theta'}{4} + (\hat{x}_3 + \hat{x}_4) \log \frac{1-\theta'}{4} + E\left(\log \frac{197!}{\Pi_{j=1}^5 x_j!}\right). \tag{5}$$

Now suppose that one seeks the MPLE $\hat{\theta}$ which maximizes $L(\theta) - \lambda(\theta - 1/2)^2$. Differentiation of $Q(\theta'|\theta) - \lambda(\theta' - 1/2)^2$ with respect to θ' yields

$$\frac{\hat{x}_2 + \hat{x}_5}{\theta'} - \frac{\hat{x}_3 + \hat{x}_4}{1-\theta'} - \lambda(2\theta' - 1) = 0. \tag{6}$$

Note that the last term of $Q(\theta'|\theta)$ in (5) is independent of θ' and so does not contribute when differentiating $Q(\theta'|\theta)$ with respect to θ'.

The EM algorithm therefore takes the form

$$\begin{cases} \hat{x}_2 = y_1\theta^n/(2+\theta^n), \\ (\hat{x}_2 + y_4)/\theta^{n+1} - (y_2 + y_3)/(1-\theta^{n+1}) - \lambda(2\theta^{n+1} - 1) = 0, n = 0, 1, 2, \ldots, \end{cases} \tag{7}$$

which yields a cubic equation for updating θ. The OSL method uses the *current* value of θ in the last term of (6), so the iteration scheme becomes

$$(8)\quad \begin{cases} \hat{x}_2 = y_1\theta^n/(2+\theta^n), \\ (\hat{x}_2 + y_4)/\theta^{n+1} - (y_2+y_3)/(1-\theta^{n+1}) - \lambda(2\theta^n - 1) = 0, n = 0, 1, 2, \ldots, \end{cases}$$

which yields a quadratic equation for updating θ.

3. The two-step-late iteration. In some ways, one can view the OSL method as being the first step of an operator splitting strategy for the maximization of (2). If this point of view is adopted, then one is led naturally to seek a more involved iterative strategy than simply applying OSL alone. In fact, the TSL method, defined below, amounts to a simple modification of the second of two steps of OSL.

Consider the following TSL iteration:

$$(9)\qquad D_{10}Q(\boldsymbol{\theta}^{(n+1/2)}|\boldsymbol{\theta}^n) - \lambda DJ(\boldsymbol{\theta}^n) = 0,$$

$$(10)\qquad D_{10}Q(\boldsymbol{\theta}^{n+1}|\boldsymbol{\theta}^{(n+1/2)}) - \lambda DJ(\boldsymbol{\theta}^n) = 0, \qquad n = 0, 1, 2, \ldots.$$

Let N denote the TSL map $N : \boldsymbol{\theta}^n \mapsto \boldsymbol{\theta}^{n+1}$; i.e., $\boldsymbol{\theta}^{n+1} = N(\boldsymbol{\theta}^n)$, where $\boldsymbol{\theta}^{n+1}$ is defined implicitly by (9) and (10). Differentiating (10) with respect to $\boldsymbol{\theta}^n$ yields

$$D_{11}Q(\boldsymbol{\theta}^{n+1}|\boldsymbol{\theta}^{(n+1/2)})DM(\boldsymbol{\theta}^n) + D_{20}Q(\boldsymbol{\theta}^{n+1}|\boldsymbol{\theta}^{(n+1/2)})DN(\boldsymbol{\theta}^n) - \lambda D_2 J(\boldsymbol{\theta}^n) = 0.$$

If $\{\boldsymbol{\theta}^n\}_{n\geq 0}$ converges to $\hat{\boldsymbol{\theta}}$, then by continuity

$$D_{11}Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})DM(\hat{\boldsymbol{\theta}}) + D_{20}Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})DN(\hat{\boldsymbol{\theta}}) - \lambda D_2 J(\hat{\boldsymbol{\theta}}) = 0$$

and consequently

$$\begin{aligned} DN(\hat{\boldsymbol{\theta}}) &= [-D_{20}Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})]^{-1}[D_{11}Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})DM(\hat{\boldsymbol{\theta}}) - \lambda D_2 J(\hat{\boldsymbol{\theta}})] \\ &= (B+C)^{-1}(C(B+C)^{-1}(C-\lambda K) - \lambda K). \end{aligned}$$

Note. TSL is just one example of a procedure resembling a two-point iteration scheme [13, pp. 336–339] for the solution of MPLE problems of the form (2).

Comparing the OSL and TSL algorithms, one can easily see from (4), (9), and (10) that

1. TSL is as practical and as easily implemented as OSL,
2. the limit points of the OSL and TSL algorithms satisfy the equation

$$(11)\qquad DL(\boldsymbol{\theta}) - \lambda DJ(\boldsymbol{\theta}) = 0.$$

For some problems, (11) has a unique solution for a large class of $J(\boldsymbol{\theta})$ and any suitably small $\lambda > 0$ [16]. In this situation, if the OSL and TSL algorithms converge, they converge to the same point. The general convergence of the TSL algorithm, though, like the OSL algorithm, remains an open question.

In what follows, we assume that for any suitably small positive λ both OSL and TSL converge and that (11) has a unique solution. Hence, when OSL and TSL converge, they do so to this solution.

Remark 2. There are two oversights in [7]. First, the dependence of B, C, and K on λ is not mentioned; second, the proof of the proposition doesn't make sense unless both the EM algorithm without penalty (i.e., $\lambda = 0$) and the OSL algorithm

(for $\lambda > 0$) converge to the same point. This is typically not the case because the solutions of (11) for $\lambda = 0$ and $\lambda > 0$ are usually different.

Since TSL corresponds to one step of OSL followed by a simple modification of OSL before the process is repeated, it is a natural consequence that (cf. equations (9) and (10)) one TSL step corresponds computationally to two steps of OSL. Thus, in order to compare the asymptotic rates of convergence of OSL and TSL, it is necessary to compare the spectral radii, respectively, of $(DM)^2$ and DN or, equivalently, of

$$G_1 = \Delta^2 \quad \text{and} \quad G_2 = (B + C)^{-1}(C\Delta - \lambda K),$$

where $\Delta = (B + C)^{-1}(C - \lambda K)$ corresponds to $DM(\hat{\boldsymbol{\theta}})$.

In essence, the purpose of this paper is an investigation of the following proposition.

PROPOSITION. *Under suitable conditions on B, C, K, and λ,*

$$\rho(G_2) < \rho(G_1), \tag{12}$$

where $\rho(G_i)$ denotes the spectral radius of $G_i, i = 1, 2$.

Below, we prove three separate theorems which establish three independent classes of conditions on B, C, and K which guarantee that (12) holds. All three theorems require $\lambda > 0$ to be suitably small; B, C, and K to be symmetric and positive definite for nonnegative and small λ; and the matrices $(B + C)^{-1}$, C, and K to be continuous at $\lambda = 0$. In addition, the first requires that $(B + C)^{-1}C|_{\lambda=0}$ has a simple largest eigenvalue; the second requires that either spectral norm $\|B^{-1}C\|_2$ or $\|C^{-1}B\|_2$ be suitably small; while the third requires that B, C, and K commute with each other.

Remark 3. Observe that, trivially, $\rho(G_2) = \rho(G_1)$ when $\lambda = 0$ because then $G_1(0) = G_2(0) = ((B + C)^{-1}C)^2|_{\lambda=0}$. All the results below make use of a property of the matrix $(B + C)^{-1}C|_{\lambda=0}$ which is just the linearization of the nonlinear map determined by the OSL algorithm applied without penalty (i.e., $\lambda = 0$).

THEOREM 1. *For small nonnegative λ, if*

(i) *B, C, and K are symmetric and positive definite,*

(ii) *$(B + C)^{-1}$, C, and K are continuous at $\lambda = 0$ with respect to λ,*

(iii) *$(B + C)^{-1}C|_{\lambda=0}$ has a simple largest eigenvalue,*

then $\rho(G_2) < \rho(G_1)$ for sufficiently small positive λ.

Proof. Since $(B + C)^{-1}$ is symmetric and positive definite, it has a Cholesky decomposition

$$(B + C)^{-1} = LL^T, \tag{13}$$

where L is a real matrix. Hence, G_1 is similar to $L^T(C - \lambda K)\Delta L$; i.e., $L^T(C - \lambda K)(B + C)^{-1}(C - \lambda K)L$, which is symmetric and positive definite when $C - \lambda K$ is invertible, and therefore this matrix and G_1 have the same eigenvalues, all of which are real and positive.

From [7, p. 446], all eigenvalues of $(B + C)^{-1}C$ are real and positive and less than one. Suppose $(B + C)^{-1}|_{\lambda=0}$ has the largest eigenvalue $\alpha < 1$, and let the corresponding eigenvector be $\boldsymbol{y}$; i.e., $(B+C)^{-1}C\boldsymbol{y} = \alpha\boldsymbol{y}$. Consequently, $(1-\alpha)C\boldsymbol{y} = \alpha B\boldsymbol{y}$, and therefore

$$(B + C)^{-1}B\boldsymbol{y} = \frac{1-\alpha}{\alpha}(B + C)^{-1}C\boldsymbol{y} = \frac{1-\alpha}{\alpha} \cdot \alpha\boldsymbol{y} = (1 - \alpha)\boldsymbol{y}. \tag{14}$$

The largest eigenvalue and the corresponding eigenvector of the matrix $(B + C)^{-1}C(B + C)^{-1}C|_{\lambda=0}(= G_1(0) = G_2(0))$ are, respectively, α^2 and $\boldsymbol{y}$. Let the eigenvalues of largest modulus of the OSL and TSL matrices G_1 and G_2 be β_1 and β_2, and let the corresponding eigenvectors be $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, respectively. Clearly,

$$\rho(G_1) = \beta_1 \quad \text{and} \quad \rho(G_2) = |\beta_2|. \tag{15}$$

It follows from [9] that

$$\begin{aligned} &\beta_1 = \alpha^2 + o(1), \quad \boldsymbol{y}_1 = \boldsymbol{y} + \boldsymbol{o}(1), \\ &\beta_2 = \alpha^2 + o(1), \quad \boldsymbol{y}_2 = \boldsymbol{y} + \boldsymbol{o}(1) \quad \text{as } \lambda \to 0. \end{aligned} \tag{16}$$

Here, $\lim_{\lambda\to 0} o(1) = 0$ and $\lim_{\lambda\to 0} \|\boldsymbol{o}(1)\| = 0$. β_1 is real and positive.

Because

$$\begin{aligned} G_2 &= (B + C)^{-1}(C\Delta - \lambda K\Delta + \lambda(B + C)^{-1}K\Delta - \lambda K) \\ &= \Delta^2 - \lambda(B + C)^{-1}K(I - \Delta), \end{aligned}$$

it follows that

$$\Delta^2\boldsymbol{y}_2 - \lambda(B + C)^{-1}K(I - \Delta)\boldsymbol{y}_2 = \beta_2\boldsymbol{y}_2. \tag{17}$$

Let $\boldsymbol{x} = L^{-1}\boldsymbol{y}_2$ (i.e., $\boldsymbol{y}_2 = L\boldsymbol{x}$). Using (13) and premultiplying (17) by $\boldsymbol{x}^*L^{-1}$ yields

$$\boldsymbol{x}^*L^T(C - \lambda K)\Delta L\boldsymbol{x} - \beta_2\boldsymbol{x}^*\boldsymbol{x} = \lambda\boldsymbol{y}_2^*K(I - \Delta)\boldsymbol{y}_2. \tag{18}$$

Expanding the right-hand side of (18) gives

$$\begin{aligned} \boldsymbol{y}_2^*K(I - \Delta)\boldsymbol{y}_2 &= \boldsymbol{y}_2^*K(B + C)^{-1}(B + \lambda K)\boldsymbol{y}_2 \\ &= \boldsymbol{y}_2^*K(B + C)^{-1}B\boldsymbol{y}_2 + \lambda\boldsymbol{y}_2^*K(B + C)^{-1}K\boldsymbol{y}_2. \end{aligned} \tag{19}$$

Obviously, $\lambda\boldsymbol{y}_2^*K(B + C)^{-1}K\boldsymbol{y}_2 > 0$ because $(B + C)^{-1}$ and K are symmetric and positive definite. Using (14) and (16) in the first term on the right-hand side of (19) yields

$$\begin{aligned} &\boldsymbol{y}_2^*K(B + C)^{-1}B\boldsymbol{y}_2 \\ &\quad = (\boldsymbol{y} + \boldsymbol{o}(1))^*K(B + C)^{-1}B(\boldsymbol{y} + \boldsymbol{o}(1)) \\ &\quad = (1 - \alpha)\boldsymbol{y}^*K\boldsymbol{y} + (1 - \alpha)\boldsymbol{o}(1)^*K\boldsymbol{y} \\ &\qquad + \boldsymbol{y}^*K(B + C)^{-1}B\boldsymbol{o}(1) + o(1)^*K(B + C)^{-1}B\boldsymbol{o}(1), \end{aligned}$$

and hence from (18)

$$\begin{aligned} &\boldsymbol{x}^*L^T(C - \lambda K)\Delta L\boldsymbol{x} - \beta_2\boldsymbol{x}^*\boldsymbol{x} \\ &= \lambda(1 - \alpha)\boldsymbol{y}^*K\boldsymbol{y} + \lambda^2\boldsymbol{y}_2^*K(B + C)^{-1}K\boldsymbol{y}_2 + o(\lambda) \quad \text{as } \lambda \to 0. \end{aligned} \tag{20}$$

Assume that $\beta_2 = \beta_{21} + i\beta_{22}$, where β_{21} and β_{22} are real. Since $L^T(C - \lambda K)\Delta L$ is symmetric positive definite, $\boldsymbol{x}^*L^T(C - \lambda K)\Delta L\boldsymbol{x}$ is real and positive. Furthermore,

$$\lambda(1 - \alpha)\boldsymbol{y}^*K\boldsymbol{y} + \lambda^2\boldsymbol{y}_2^*K(B + C)^{-1}K\boldsymbol{y}_2 (= O(\lambda))$$

is real and positive (for $\lambda > 0$), and hence (20) can be written as

$$[\boldsymbol{x}^*L^T(C - \lambda K)\Delta L\boldsymbol{x} - \beta_{21}\boldsymbol{x}^*\boldsymbol{x}] - i\beta_{22}\boldsymbol{x}^*\boldsymbol{x} = O(\lambda) + io(\lambda), \tag{21}$$

where $O(\lambda) > 0$, provided λ remains positive. Now from (21) β_{21}, the real part of β_2, which is positive, is less than β_1, and the difference $\beta_1 - \beta_{21}$ is $O(\lambda)$. The imaginary part β_{22} of β_2 is $o(\lambda)$, while that of β_1 is 0. Therefore, for sufficiently small positive λ,

$$|\beta_2| < \beta_1,$$

which verifies that the spectral radius of G_2 is less than that of G_1 for sufficiently small positive λ. □

The following lemma is required for a proof of Theorem 2 and assists in further identifying alternative conditions under which the proposition holds.

LEMMA 1. *If B, C, and K are symmetric and positive definite and either*

(i) *the spectral norm of the matrix $B^{-1}C$ satisfies $q := \|B^{-1}C\|_2 < 1$ or*

(ii) *the spectral norm of the matrix $C^{-1}B$ satisfies $q := \|C^{-1}B\|_2 < 1$,*

then for any unit vector $\boldsymbol{x}$, $\boldsymbol{x}^*K(I-\Delta)\boldsymbol{x}$ is a complex number which correspondingly satisfies

(i)
$$\begin{aligned} &\mathrm{Re}(\boldsymbol{x}^*K(I-\Delta)\boldsymbol{x}) > \xi_{\min} + \lambda\mu_{\min} - O(q), \\ &|\mathrm{Im}(\boldsymbol{x}^*K(I-\Delta)\boldsymbol{x})| < O(q) \quad \text{as } q \to 0, \end{aligned}$$

(ii)
$$\begin{aligned} &\mathrm{Re}(\boldsymbol{x}^*K(I-\Delta)\boldsymbol{x}) > \lambda\mu_{\min} - O(q), \\ &|\mathrm{Im}(\boldsymbol{x}^*K(I-\Delta)\boldsymbol{x})| < O(q) \quad \text{as } q \to 0, \end{aligned}$$

where $\xi_{\min}$ and $\mu_{\min}$ are positive and denote, respectively, the smallest eigenvalues of K and $K(B+C)^{-1}K$, and $O(q)$ is positive.

Proof. Clearly

$$\begin{aligned} K(I-\Delta) &= K(I-(B+C)^{-1}(C-\lambda K)) \\ &= K(B+C)^{-1}(B+\lambda K) \\ &= K(B+C)^{-1}B + \lambda K(B+C)^{-1}K. \end{aligned}$$

(i) When $\|B^{-1}C\| = q < 1$, then [11]

$$\begin{aligned} (B+C)^{-1}B &= (B^{-1}(B+C))^{-1} \\ &= (I+B^{-1}C)^{-1} \\ &= I - B^{-1}C + B^{-1}CB^{-1}C - \cdots. \end{aligned}$$

Thus, for any vector $\boldsymbol{x}$

$$\begin{aligned} &\boldsymbol{x}^*K(I-\Delta)\boldsymbol{x} \\ &= \boldsymbol{x}^*K\boldsymbol{x} + \boldsymbol{x}^*K(-B^{-1}C + B^{-1}CB^{-1}C - \cdots)\boldsymbol{x} + \lambda\boldsymbol{x}^*K(B+C)^{-1}K\boldsymbol{x}. \end{aligned}$$

If $\boldsymbol{x}$ is a unit vector, then

$$\begin{aligned} &|\boldsymbol{x}^*K(-B^{-1}C + B^{-1}CB^{-1}C - \cdots)\boldsymbol{x}| \\ &\le \|K\|(\|B^{-1}C\| + \|B^{-1}C\|^2 + \|B^{-1}C\|^3 + \cdots) \\ &= \frac{q\xi_{\max}}{1-q} = O(q) \quad \text{as } q \to 0, \end{aligned}$$

where $\xi_{\max}$, which is the largest eigenvalue of K, is also its spectral norm.

On the other hand,

$$\begin{aligned} K(I-\Delta) &= K(I-(B+C)^{-1}(C-\lambda K)) \\ &= K - K(B+C)^{-1}C + \lambda K(B+C)^{-1}K. \end{aligned}$$

(ii) When $\|C^{-1}B\| = q < 1$, then [11]

$$\begin{aligned} (B+C)^{-1}C &= (C^{-1}(B+C))^{-1} \\ &= (I+C^{-1}B)^{-1} \\ &= I - C^{-1}B + C^{-1}BC^{-1}B - \cdots, \end{aligned}$$

and so

$$\begin{aligned} &\boldsymbol{x}^* K(I-\Delta)\boldsymbol{x} \\ &= \boldsymbol{x}^* K(-C^{-1}B + C^{-1}BC^{-1}B - \cdots)\boldsymbol{x} + \lambda \boldsymbol{x}^* K(B+C)^{-1}K\boldsymbol{x}. \end{aligned}$$

If $\boldsymbol{x}$ is a unit vector, then

$$\begin{aligned} &|\boldsymbol{x}^* K(-C^{-1}B + C^{-1}BC^{-1}B - \cdots)\boldsymbol{x}| \\ &\le \|K\|(\|C^{-1}B\| + \|C^{-1}B\|^2 + \|C^{-1}B\|^3 + \cdots) \\ &= \frac{q\xi_{\max}}{1-q} = O(q) \quad \text{as } q \to 0. \end{aligned}$$

Because $\boldsymbol{x}^* K\boldsymbol{x}$ and $\lambda \boldsymbol{x}^* K(B+C)^{-1}K\boldsymbol{x}$ are real and

$$\begin{aligned} &\boldsymbol{x}^* K\boldsymbol{x} \ge \xi_{\min}, \\ &\lambda \boldsymbol{x}^* K(B+C)^{-1}K\boldsymbol{x} \ge \lambda\mu_{\min} \quad \text{and} \\ &|\boldsymbol{x}^* K(-B^{-1}C + B^{-1}CB^{-1}C - \cdots)\boldsymbol{x}| < O(q). \end{aligned}$$

Then Lemma 1 is proved. □

We are now in a position to state and prove our second theorem. Lemma 1 provides one set of the aforementioned alternative conditions under which the proposition holds.

THEOREM 2. *If*

(i) *B, C, and K are symmetric and positive definite,*

(ii) *the spectral norm $\|B^{-1}C\|_2$ (or $\|C^{-1}B\|_2$) is sufficiently small in a neighborhood of $\lambda = 0$, $\lambda > 0$,*

(iii) *$(B+C)^{-1}$, C, and K are continuous at $\lambda = 0$ with respect to λ,*

then $\rho(G_2) < \rho(G_1)$ for sufficiently small positive λ.

Proof. Suppose γ is an eigenvalue of G_2 and $\boldsymbol{z}$ is the corresponding eigenvector such that $\|\boldsymbol{z}\| = 1$. Let $\boldsymbol{x} = L^{-1}\boldsymbol{z}$, where L is as given in (13). Then, proceeding in a manner similar to the proof of Theorem 1, we get

$$\boldsymbol{x}^* L^T(C-\lambda K)\Delta L\boldsymbol{x} - \gamma \boldsymbol{x}^*\boldsymbol{x} = \lambda \boldsymbol{z}^* K(I-\Delta)\boldsymbol{z}. \tag{22}$$

From Lemma 1, the real part of γ is less than the spectral radius of G_1, and the difference is greater than

$$\frac{\lambda}{\boldsymbol{x}^*\boldsymbol{x}}(\xi_{\min} + \lambda\mu_{\min} - O(q)) \qquad \left(or\ \frac{\lambda}{\boldsymbol{x}^*\boldsymbol{x}}(\lambda\mu_{\min} - O(q))\right),$$

while the difference of imaginary parts is less than $\lambda O(q)$. By the continuity of γ with respect to λ [9], when λ is sufficiently small and positive, the real part of γ is positive, because when $\lambda \to 0$, γ approximates one eigenvalue of $((B+C)^{-1}C)^2|_{\lambda=0}$, which is positive. Therefore, for sufficiently small q, the absolute value of γ is less than the spectral radius of G_1. This is true for each eigenvalue of G_2, so $\rho(G_2) < \rho(G_1)$. □

Finally, we give a third result in which TSL is superior to OSL. However, in this case the assumed conditions may be more difficult to satisfy in practice for higher dimensional problems.

THEOREM 3. *If*

(i) *B, C, and K are symmetric and positive definite matrices which commute with each other (i.e., $BC = CB, BK = KB, CK = KC$),*

(ii) *$(B+C)^{-1}$, C, and K are continuous at $\lambda = 0$ with respect to λ,*

then for sufficiently small positive λ, $\rho(G_2) < \rho(G_1)$.

Proof. Assume the same notation as in the proof of Theorem 1. From [7, p. 446] as well as above, all eigenvalues of $(B+C)^{-1}C$ are real and positive. Under the commutativity condition, the TSL matrix G_2 is symmetric, so all of its eigenvalues are real. Furthermore, when $\lambda = 0$, G_2 reduces to be $((B+C)^{-1}C)^2|_{\lambda=0}$, for which the largest eigenvalue is α^2. Hence, for sufficiently small λ and because of the continuity of the eigenvalues of G_2 with respect to λ, β_2 is positive.

It easily follows from the commutativity condition and the positive definiteness of B, C, and K that $K(B+C)^{-1}B$ is positive definite. Hence, for any nonzero vector $\boldsymbol{u}$, $\boldsymbol{u}^* K(B+C)^{-1}B\boldsymbol{u} > 0$, and, therefore, from (19)

$$\boldsymbol{y}_2^* K(I-\Delta)\boldsymbol{y}_2 > 0.$$

In this case for positive λ, (18) becomes

$$\boldsymbol{x}^* L^T (C - \lambda K)\Delta L \boldsymbol{x} - \beta_2 \boldsymbol{x}^* \boldsymbol{x} = \lambda \boldsymbol{y}_2^* K(I-\Delta)\boldsymbol{y}_2 > 0. \tag{23}$$

Hence $\beta_2 < \rho(L^T(C-\lambda K)\Delta L)$; i.e., $\beta_2 < \rho(G_1)$. Since $\rho(G_2) = \beta_2$, then $\rho(G_2) < \rho(G_1)$ for sufficiently small positive λ. □

Theorems 1, 2, and 3 establish conditions which guarantee that the TSL algorithm asymptotically converges faster than the OSL algorithm and give different verifications of the proposition.

Remark 4. The fact that in all three theorems λ must be sufficiently small does not pose a problem since λ plays the role of the regularization parameter which, by its very nature, must be kept appropriately small in order not to generate oversmoothed approximations $\hat{\boldsymbol{\theta}}$.

4. A numerical example. To illustrate the practical applicability and usefulness of the above results, we consider the same test problem as that used in [4] and [7], namely, a multinomial sample for which the parameter space is one dimensional and Theorem 3 automatically holds. Results corresponding to those given below for this example have also been obtained for ridge regression [17].

Let the setting be the same as that in section 2. The EM and OSL algorithms for this problem have been derived in section 2 as equations (7) and (8), respectively.

The TSL iteration scheme for this problem takes the form

$$\begin{cases} x_2^{(1/2)} = y_1\theta^n/(2+\theta^n), \\ (x_2^{(1/2)} + y_4)/\theta^{(n+1/2)} - (y_2+y_3)/(1-\theta^{(n+1/2)}) - \lambda(2\theta^n - 1) = 0, \\ \hat{x}_2 = y_1\theta^{(n+1/2)}/(2+\theta^{(n+1/2)}), \\ (\hat{x}_2 + y_4)/\theta^{n+1} - (y_2+y_3)/(1-\theta^{n+1}) - \lambda(2\theta^n - 1) = 0, \quad n = 0, 1, 2, \ldots. \end{cases}$$

TABLE 1
Convergence in a multinomial example.

λ	$\hat{\boldsymbol{\theta}}$	OSL^2	**TSL**
0.001	0.6268214	0.01765	0.01761
0.005	0.6268187	0.01762	0.01760
0.01	0.6268153	0.01762	0.01758
0.02	0.6268086	0.01761	0.01753
0.04	0.6267951	0.01759	0.01743
0.05	0.6267884	0.01757	0.01738
0.1	0.6267548	0.0175	0.0171
0.2	0.6266878	0.0174	0.0166
0.3	0.6266207	0.0173	0.0161
0.4	0.6265537	0.0172	0.0156
0.5	0.6264863	0.0171	0.0151
1.0	0.6261530	0.0165	0.0125
2.0	0.6254897	0.0155	0.0074
3.0	0.6248319	0.0144	0.0023
4.0	0.6241797	0.0134	0.0029
5.0	0.6235331	0.0125	0.0081
6.0	0.6228920	0.0115	0.0133

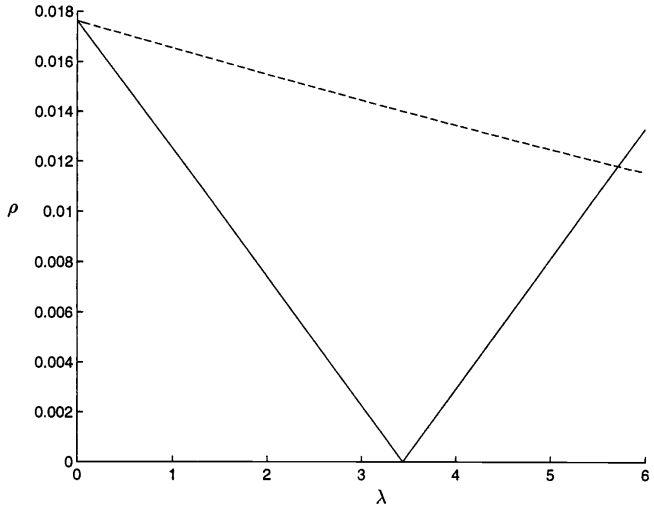


FIG. 1. *Approximate rate of convergence ρ plotted against λ for the multinomial sample: dashed line OSL^2, solid line TSL.*

Both OSL and TSL converge to the same solution even for different initial estimates θ^0. Numerical experimentation confirms that the rates of convergence of OSL and TSL vary with λ and that for sufficiently small positive λ, TSL is superior to OSL. The numerical rates of convergence which demonstrate this are illustrated in Table 1 and Figure 1, where the notation OSL^2 is used to denote two steps of OSL, the computational equivalent of one step of TSL. These data were obtained using $\theta^0 = 0.4$. The numbers in the third and fourth columns of Table 1 are the quantity $\rho = |\theta^N - \theta^{N-1}|/|\theta^{N-1} - \theta^{N-2}|$ for the last iterate before convergence to $\hat{\theta}$. Convergence here is defined by $|\theta^N - \theta^{N-1}| \leq 10^{-8}$.

From Figure 1 it is clear that for this example, the size of λ for which the better convergence of TSL is achieved exceeds those values likely to be useful in practice (cf. Remark 4).

Acknowledgment. The authors wish to thank Mark Westcott for providing valuable insights into the EM methodology.

REFERENCES

[1] R. S. Anderssen, G. A. Latham, and M. Westcott, *Statistical methodology for inverse problems*, Math. Comput. Modelling, 22 (1995), pp. 1–5.
[2] R. A. Boyles, *On the convergence of the EM algorithm*, J. Roy. Statist. Soc. B, 45 (1983), pp. 47–50.
[3] I. Csiszàr and G. Tusnàdy, *Information geometry and alternating minimization procedures*, Statistics & Decisions, Suppl., 1 (1984), pp. 205–237.
[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm (with discussion)*, J. Roy. Statist. Soc. Ser. B, 39 (1977), pp. 1–38.
[5] R. Glowinski, *Splitting methods for the numerical solution of the incompressible Navier–Stokes equations*, in Vistas in Applied Mathematics, A. V. Balakrishnan, A. A. Dorodnitsyn, and J. L. Lions, eds., Optimization Software, New York, 1986, pp. 57–95.
[6] P. J. Green, *Bayesian reconstructions from emission tomography data using a modified EM algorithm*, IEEE Trans. Med. Imaging, 9 (1990), pp. 84–93.
[7] P. J. Green, *On use of the EM algorithm for penalized likelihood estimation*, J. Roy. Statist. Soc. Ser. B, 52 (1990), pp. 443–452.
[8] C. W. Groetsch, *Inverse Problems in the Mathematical Sciences*, Vieweg Mathematics for Scientists and Engineers, Vieweg, Wiesbaden, 1993.
[9] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.
[10] A. I. Khuri, *Advanced Calculus With Applications in Statistics*, John Wiley, New York, 1993.
[11] P. Lancaster, *Theory of Matrices*, Academic Press, New York, 1969.
[12] K. Lange, *Convergence of EM image reconstruction algorithms with Gibbs smoothing*, IEEE Trans. Med. Imaging, 9 (1990), pp. 439–446.
[13] A. Ralston, *A first course in numerical analysis*, in International Series in Pure and Applied Mathematics, McGraw–Hill, New York, 1965.
[14] L. A. Shepp and Y. Vardi, *Maximum likelihood reconstruction in positron emission tomography*, IEEE Trans. Med. Imaging, 1 (1982), pp. 113–122.
[15] C. F. J. Wu, *On the convergence properties of the EM algorithm*, Ann. Statist., 11 (1983), pp. 95–103.
[16] S. Yu, G. A. Latham, and R. S. Anderssen, *Stabilizing properties of maximum penalized likelihood estimation for additive Poisson regression*, Inverse Problems, 10 (1994), pp. 1199–1209.
[17] S. Yu, G. A. Latham, and R. S. Anderssen, *Two-Stage Iteration for Maximum Penalized Likelihood Estimation*, Centre for Mathematics and its Applications, Australian National University, Canberra, Australia, Research Report CMA-MR44-93.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, pp. 360–368, April 1997

008

CONVERGENCE OF POLYNOMIALLY BOUNDED SEMIGROUPS OF MATRICES*

LEONID GURVITS† AND LEIBA RODMAN‡

Abstract. It is proved that for polynomially bounded sets of matrices the notions of pointwise convergence and uniform convergence coincide. This result is also proved for certain sets of nonlinear maps on finite-dimensional real or complex vector spaces.

Key words. uniform convergence, pointwise convergence, infinite products, matrix semigroups

AMS subject classifications. 15A30, 15A99

PII. S089547989528939X

1. Introduction. Let $\mathcal{A}$ be a set of $n \times n$ matrices with complex or real entries. Various notions concerning convergence of infinite products of matrices in $\mathcal{A}$ have been studied extensively in the literature. We mention [DL2, BW, DL1], where right-convergent product sets are studied (such sets appear in many applications, for example, in constructing wavelets with compact support; see the bibliography in [DL1]). Various notions of stability of discrete linear inclusions lead to a study of infinite products of matrices and their convergence (see [G]). We also mention [S, SU], where a notion of convergence of finite matrix products is studied; the type of convergence there is motivated by the theory of multimodal linear control systems.

In this paper we mainly study pointwise convergence and uniformly convergent sets. The definitions of these notions of convergence will now be given. We set up notation and conventions first. Let $F = \mathbf{R}$ or $F = \mathbf{C}$. It will be convenient to represent the set $\mathcal{A}$ as indexed by some index set K; thus, we write $\mathcal{A} = \{A_i : i \in K\}$, where each A_i is an $n \times n$ matrix with entries in F. A *word* w of length k is by definition a function $w : \{1, \dots, k\} \to K$; we denote by $|w|$ the length of the word w. For a given word w, let A_w be the left product $A_{w(k)}A_{w(k-1)} \cdots A_{w(1)}$, where $k = |w|$. Sometimes we consider words of infinite length; i.e., functions $w : \{1, 2, \dots\} \to K$; for such a word w we denote by $w^{(k)}$ the restriction of w to the finite set $\{1, 2, \dots, k\}$. Thus $w^{(k)}$ has length k. A set $\mathcal{A} = \{A_i : i \in K\}$ is called *pointwise convergent* (more precisely, pointwise convergent to zero) if for every $x \in F^n$ there is a word w of infinite length (which may depend on x) such that

$$\lim_{k\to\infty} (A_{w^{(k)}} x) = 0.$$

The set $\mathcal{A}$ is called *uniformly convergent* if there is a word w of infinite length such that

$$\lim_{k\to\infty} (A_{w^{(k)}} x) = 0 \tag{1.1}$$

for all $x \in F^n$. Clearly, every uniformly convergent set is pointwise convergent. It is well known that the converse is generally false (examples to illustrate this fact, as well

* Received by the editors July 24, 1995; accepted for publication by R. Brualdi May 15, 1996. http://www.siam.org/journals/simax/18-2/28939.html

† NEC Research Institute, 4 Independence Way, Princeton, NJ 08540 (gurvits@research.nj.nec.com).

‡ Department of Mathematics, The College of William and Mary, Williamsburg, VA 23187-8795 (lxrodm@math.wm.edu). This author was partially supported by NSF grant DMS-9500924.

as some basic information on pointwise convergent and uniformly convergent sets, are given in section 2). The main result of this paper is the following theorem.

THEOREM 1.1. *Assume that the set $\mathcal{A} = \{A_i : i \in K\}$ of $n \times n$ matrices over F generates a polynomially bounded semigroup*; *i.e.,*

$$\sup_{|w|=k} \|A_w\| \le Ck^p, \quad k = 1, 2, \dots, \tag{1.2}$$

where the positive constants C and p are independent of k and where some norm $\|\cdot\|$ on the algebra of all $n \times n$ matrices over F is used (the property of being polynomially bounded is obviously independent of $\|\cdot\|$). Then $\mathcal{A}$ is pointwise convergent if and only if $\mathcal{A}$ is uniformly convergent.

Theorem 1.1 will be proved in section 3. The main ingredient in the proof is a probabilistic argument (this type of argument originates in [G]).

In section 4 we present various generalizations and extensions of Theorem 1.1. In particular, it turns out that the linearity of transformations $x \mapsto Ax$, $x \in F^n$, where $A \in \mathcal{A}$ is fixed, plays a secondary role in Theorem 1.1; in fact, this result (assuming $\mathcal{A}$ generates a bounded semigroup) is extended in section 4 to a large class of nonlinear maps. Also, an abstract setting in which a result analogous to Theorem 1.1 is proved is presented in the same section 4.

We conclude the introduction with an illustrative example where Theorem 1.1 is applied.

Example 1.1. Let A be an $n \times n$ matrix with entries in F such that the singular values of A do not exceed 1, and A has an eigenvalue $\lambda \in \mathbf{C}$ with $|\lambda| < 1$. Let $\{U_i\}_{i \in K}$, $U_i \in F^{n \times n}$, be a semigroup of unitary matrices which is *almost transitive*; i.e., for every $x \in F^n$ having the Euclidean norm equal to 1 and every $\epsilon > 0$ there is $i \in K$ such that $\|U_i e_1 - x\| < \epsilon$. Here $e_1 = (1, 0, \dots, 0)^T \in F^n$. Then the set $\mathcal{A} = \{U_i : i \in K\} \cup \{A\}$ is uniformly convergent. Indeed, the almost transitive property of $\{U_i\}_{i \in K}$ and the existence of $\lambda \in \sigma(A)$ with $|\lambda| < 1$ guarantee that $\mathcal{A}$ is pointwise convergent (see Proposition 2.2). Clearly, the set $\mathcal{A}$ generates a bounded semigroup; in fact, the largest singular value of any (finite) product of matrices in $\mathcal{A}$ does not exceed 1. By Theorem 1.1, $\mathcal{A}$ is uniformly convergent. ∎

2. Pointwise and uniform convergence: Preliminaries. As in the previous section, $\mathcal{A} = \{A_i : i \in K\}$ is a set of $n \times n$ matrices with entries in F, where either $F = \mathbf{R}$ or $F = \mathbf{C}$.

PROPOSITION 2.1. *$\mathcal{A}$ is uniformly convergent if and only if there exists a finite product of matrices in $\mathcal{A}$ whose spectral radius is less than* 1.

This is Theorem 3.5 in [SU].

Using the well-known fact that, for a single $n \times n$ matrix X, the condition $\lim_{k \to \infty}(X^k x) = 0$ for all $x \in F^n$ (or, equivalently, $\lim_{k \to \infty} X^k = 0$) is equivalent to the spectral radius of X being less than 1, Proposition 2.1 can be reformulated as follows. We say that a set $\mathcal{A}$ is *periodically uniformly convergent* if (1.1) holds for some periodic infinite word w. It follows from Proposition 2.1 that a set $\mathcal{A}$ is uniformly convergent if and only if it is periodically uniformly convergent.

A set $\mathcal{A}$ is called *precontractive* with respect to a norm $\|\cdot\|$ in F^n if for every $x \in F^n$, $x \neq 0$, there is a word w such that $\|A_w x\| < \|x\|$.

PROPOSITION 2.2. *The following statements are equivalent for a set $\mathcal{A}$ of $n \times n$ matrices over F:*

(i) *$\mathcal{A}$ is pointwise convergent,*

(ii) *$\mathcal{A}$ is precontractive with respect to some norm in F^n,*

(iii) *$\mathcal{A}$ is precontractive with respect to every norm in F^n.*

For a proof, see Theorem 1 in [S] and Theorem 3.2 in [SU].

By analogy with the notion of uniform convergence, we introduce the following definition. A set $\mathcal{A}$ is called *periodically pointwise convergent* if for every $x \in F^n$ there is a word w (which depends on x) such that

$$\lim_{k\to\infty} (A_w^k x) = 0.$$

We now give two examples showing that the notions of pointwise convergence, periodically pointwise convergence, and uniform convergence are all distinct.

Example 2.1. Let $F = \mathbf{R}$, $\mathcal{A} = \{A_1, A_2\}$, where

$$A_1 = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}, \qquad A_2 = \begin{bmatrix} \gamma & 0 \\ 0 & \mu \end{bmatrix}.$$

Here ϕ is a fixed real number such that $(2\pi)^{-1}\phi$ is irrational (so A_1 is an "irrational rotation"), and γ and μ are fixed positive numbers such that $\mu < 1 < \gamma$ and $\mu\gamma = 1$. By Proposition 2.2, $\mathcal{A}$ is pointwise convergent (indeed, for every $x \in \mathbf{R}^2$, $\|x\| = 1$, there is a positive integer m such that $\|A_1^m x - (0,1)^T\| < ((1-\mu^2)(\gamma^2-1))^{\frac{1}{2}}$; then $\|A_2 A_1^m x\| < \|x\|$). On the other hand, $\mathcal{A}$ is not periodically pointwise convergent. Indeed, for every word w we have $\det A_w = 1$, and therefore there is at most one (up to a multiplication by a scalar) eigenvector of A_w corresponding to an eigenvalue having absolute value less than 1. Call this eigenvector x_w. Clearly, $\lim_{k\to\infty} A_w^k x = 0$ if and only if x is a scalar multiple of x_w. But the set of all vectors x_w such that $\|x_w\| = 1$ is at most countable (because the set of words is countable). Thus, there exists a vector y which is not a scalar multiple of any x_w. For this vector we have

$$\lim_{k\to\infty} A_w^k y \neq 0$$

for all words w; hence $\mathcal{A}$ is not periodically pointwise convergent. □

Example 2.2. Let $F = \mathbf{R}$. Fix two positive real numbers μ and γ such that $\mu < 1 < \gamma$ and $\mu\gamma > 1$. Let $\mathcal{A} = \{A_v : v \in S\}$, where S is the Euclidean unit sphere in $\mathbf{R}^2$ and A_v is the 2×2 matrix defined by the property that $A_v(xv + yv^\perp) = \mu x v + \gamma y v^\perp$; $v^\perp$ is the unit vector orthogonal to v. Since $\det A_v > 1$ for all $v \in S$, the set $\mathcal{A}$ is not uniformly convergent (see Proposition 2.1). However, for every $v \in S$ we have $A_v^k v \to 0$ as $k \to \infty$. So $\mathcal{A}$ is periodically pointwise convergent. □

In connection with Example 2.2 we note that *if $\mathcal{A} = \{A_i : i \in K\}$ is a countable (or finite) periodically pointwise convergent set, then $\mathcal{A}$ is uniformly convergent.* Indeed, let w be an arbitrary (finite) word and define

$$\Omega_w = \{x \in F^n : \lim_{k\to\infty} (A_w)^k x = 0\}.$$

Clearly, Ω_w is a subspace in F^n, and by the periodically pointwise convergence of $\mathcal{A}$ we have $F^n = \bigcup_w \Omega_w$. But the set of finite words in the at most countable alphabet K is itself countable. A countable union of subspaces coincides with F^n if and only if at least one of these subspaces itself coincides with F^n. Thus, $\Omega_w = F^n$ for some word w, which means that $\mathcal{A}$ is periodically uniformly convergent and hence $\mathcal{A}$ is uniformly convergent.

3. Proof of Theorem 1.1. The proof is based on a series of lemmas.

LEMMA 3.1. *Let $\mathcal{A}$ be a set of matrices satisfying* (1.2)*. Then the set $\mathcal{A}$ is bounded, and, up to a simultaneous similarity, all matrices in $\mathcal{A}$ have the block diagonal form*

$$A_i = \begin{bmatrix} A_{11}^{(i)} & A_{12}^{(i)} & \cdots & A_{1m}^{(i)} \\ 0 & A_{22}^{(i)} & \cdots & A_{2m}^{(i)} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & A_{mm}^{(i)} \end{bmatrix}, \quad i \in K, \tag{3.1}$$

where, for $j = 1, \ldots, m$, the jth diagonal blocks $A_{jj}^{(i)}$ have size $n_j \times n_j$ (independent of $i \in K$) and generate a bounded semigroup; *i.e.,*

$$\sup_k \sup_{|w|=k} \left\| A_{jj}^{(w(k))} A_{jj}^{(w(k-1))} \cdots A_{jj}^{(w(1))} \right\| < \infty. \tag{3.2}$$

Conversely, if $\mathcal{A} = \{A_i : i \in K\}$ is a bounded set of matrices having the block triangular form (3.1) *and satisfying* (3.2)*, then $\mathcal{A}$ generates a polynomially bounded semigroup.*

Proof. The direct part is proved in [BW, Proposition III]. We prove the converse part. Because of (3.2), there is a norm $\|\cdot\|_j$ on F^{n_j} such that the induced operator norm $\|\cdot\|_j$ on the $n_j \times n_j$ matrices satisfies

$$\left\| A_{jj}^{(i)} \right\|_j \leq 1 \text{ for all } i \in K. \tag{3.3}$$

Indeed, take

$$\|x\|_j = \max \left\{ \sup_k \sup_{|w|=k} \left\| A_{jj}^{(w(k))} \cdots A_{jj}^{(w(1))} x \right\|, \|x\| \right\},$$

where $\|\cdot\|$ is the Euclidean norm (for example). Let

$$\|x\|_* = \|x_1\|_1 + \cdots + \|x_m\|_m,$$

where

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \in F^n, \quad x_j \in F^{n_j} \quad (j = 1, \ldots, m),$$

and let $\|\cdot\|_*$ be the corresponding induced operator norm on $n \times n$ matrices. Then for every word w of length k we have $\|A_w\|_* \leq \|B^k\|$, where B is the $m \times m$ matrix whose diagonal elements are equal to 1 and whose (q, r) entry $(q < r)$ is equal to

$$\sup_{i \in K} \|A_{q,r}^{(i)}\|_{q,r}. \tag{3.4}$$

Here $\|X\|_{q,r}$ is the induced operator norm when an $n_q \times n_r$ matrix X is considered as a linear transformation from F^{n_r} (with the norm $\|\cdot\|_r$) into F^{n_q} (with the norm

$\|\cdot\|_q$). The supremum in (3.4) is finite because $\mathcal{A}$ is bounded. Now (1.2) follows with $p = m$. ∎

In general, it is not true that if the diagonal blocks of a set of diagonal matrices form uniformly convergent sets, then the whole set is uniformly convergent.

Example 3.1. Let

$$\mathcal{A} = \left\{ \begin{bmatrix} \alpha & 0 \\ 0 & \alpha^{-1} \end{bmatrix} : \alpha > 0 \right\}.$$

The set $\mathcal{A}$ is not uniformly convergent because $\det A = 1$ for every $A \in \mathcal{A}$. Nevertheless, the diagonal blocks $\{\alpha : \alpha > 0\}$ and $\{\alpha^{-1} : \alpha > 0\}$ form uniformly convergent sets of 1×1 matrices. ∎

LEMMA 3.2. *Let $\mathcal{A} = \{A_i : i \in K\}$ be given in the form* (3.1). *Assume that, for $j = 1, \dots, m$, the jth diagonal blocks $\{A_{jj}^{(i)} : i \in K\}$ generate a bounded semigroup. If the sets $\mathcal{A}_j = \{A_{jj}^{(i)} : i \in K\}$ $(j = 1, \dots, m)$ are uniformly convergent, then $\mathcal{A}$ is uniformly convergent.*

Proof. Using the induced operator norm on $n_j \times n_j$ matrices, as in the proof of Lemma 3.1, we can (and do) assume that

$$\|A_{jj}^{(i)}\|_j \le 1 \text{ for all } i \in K \text{ and } j = 1, \dots, m \tag{3.5}$$

for some induced operator norms $\|\cdot\|_j$. By Proposition 2.1, for every $j = 1, \dots, m$, there exists a word w_j such that the matrix products $A_{jj}^{(w_j)}$ satisfy

$$\|A_{jj}^{(w_j)}\|_j < 1 \qquad (j = 1, \dots, m). \tag{3.6}$$

Let $k = |w_1| + \cdots + |w_m|$ and w be the word of length k that acts as w_1 on the set $\{1, \dots, |w_1|\}$, as w_2 on the set $\{|w_1| + 1, \dots, |w_1| + |w_2|\}$, ..., as w_m on the set $\{|w_1| + \cdots + |w_{m-1}| + 1, \dots, |w_1| + \cdots + |w_m|\}$. In view of (3.5) and (3.6), we have

$$\left\| A_{jj}^{(w(k))} \cdots A_{jj}^{(w(1))} \right\|_j < 1 \text{ for } j = 1, \dots, m.$$

Thus, the spectral radius of the product

$$\left(A_{11}^{(w(k))} \oplus \cdots \oplus A_{mm}^{(w(k))} \right) \cdots \left(A_{11}^{(w(1))} \oplus \cdots \oplus A_{mm}^{(w(k))} \right)$$

is less than 1. Hence the same is true of the product $A_{w(k)} A_{w(k-1)} \cdots A_1$, and the set $\mathcal{A}$ is uniformly convergent by Proposition 2.1. ∎

LEMMA 3.3. *Let $\mathcal{A} = \{A_i : i \in K\}$ be a set of $n \times n$ matrices such that $\mathcal{A}$ generates a bounded semigroup. Then $\mathcal{A}$ is pointwise convergent if and only if $\mathcal{A}$ is uniformly convergent.*

Proof. Assume $\mathcal{A}$ is pointwise convergent. As in the proof of Lemma 3.1, assume that

$$\|A_i\| \le 1 \text{ for all } i \in K, \tag{3.7}$$

where $\|\cdot\|$ is the operator norm induced by a norm (also denoted $\|\cdot\|$) on F^n. By Proposition 2.2 for every x on the unit sphere S (with respect to the norm $\|\cdot\|$) there exists a finite product B_x of matrices in $\mathcal{A}$ such that $\|B_x x\| < \alpha_x$ for some $\alpha_x < 1$. Let U_x be an open set containing x such that $\|B_x y\| < \alpha_x \|y\|$ for every $y \in U_x$.

Using the compactness of the unit sphere S, select a finite set $U_{x_1}, \dots, U_{x_q}$ such that $\bigcup_{j=1}^{q} U_{x_j} \supseteq S$. Thus, for every $y \in S$, there exists $j \in \{1, \dots, q\}$ such that

$$\|B_{x_j} y\| \le \alpha, \tag{3.8}$$

where $0 < \alpha < 1$ is independent of y. Abbreviate $B_j = B_{x_j}$. It is sufficient to prove that the finite set $\mathcal{B} = \{B_1, \dots, B_q\}$ is uniformly convergent.

Let $\xi_1, \xi_2, \dots$ be a sequence of independent identically distributed random variables such that each ξ_j takes values in the set $\{1, \dots, q\}$ with equal probability q^{-1} of each value. We claim that for every $x \in F^n$

$$\lim(\cdots B_{\xi_m} B_{\xi_{m-1}} \cdots B_{\xi_1} x) = 0 \tag{3.9}$$

with probability 1.

We say that B_j *covers* $y \in F^n$ if $\|B_j y\| \le \alpha \|y\|$, where α is taken from (3.8). Denote by $P\{X\}$ the probability of the event X. We prove

$$P\Big\{ B_{\xi_1} \text{ does not cover } y;\ B_{\xi_2} \text{ does not cover } B_{\xi_1} y; \dots;$$

$$B_{\xi_m} \text{ does not cover } B_{\xi_{m-1}} B_{\xi_{m-2}} \cdots B_{\xi_1} y \Big\} \le ((q-1)/q)^m. \tag{3.10}$$

Inequality (3.10) will be proved by induction on m. For $m = 1$, we have

$$P\Big\{ B_{\xi_1} \text{ does not cover } y \Big\} = P\Big\{ y \ne 0 \text{ and } \xi_1 \text{ is not equal to}$$

$$\text{one of the values } j \in \{1, \dots, q\} \text{ for which } y \text{ is a scalar multiple}$$

$$\text{of a vector in } U_{x_j} \Big\} \le (q-1)/q \tag{3.11}$$

by the definition of ξ_1. Assume (3.10) is proved with m replaced by $m-1$. Denote by $S_m = S_m(y)$ the event

$$B_{\xi_1} \text{ does not cover } y; \dots; B_{\xi_m} \text{ does not cover } B_{\xi_{m-1}} B_{\xi_{m-2}} \cdots B_{\xi_1} y.$$

Then, the left-hand side of (3.10) is

$$P\{S_{m-1}\} P\{B_{\xi_m} \text{ does not cover } B_{\xi_{m-1}} \cdots B_{\xi_1} y \,|\, S_{m-1}\}.$$

Observe that under the conditions that the event S_{m-1} happened we have

$$B_{\xi_{m-1}} \cdots B_{\xi_1} y \ne 0.$$

Let $z = B_{\xi_{m-1}} \cdots B_{\xi_1} y$; the vector z is a random vector which (under the condition that S_{m-1} happened) takes a finite number of nonzero values $z_1, \dots, z_p$ with probabilities $\alpha_1, \dots, \alpha_p$, respectively. Thus

$$P\Big\{ B_{\xi_m} \text{ does not cover } B_{\xi_{m-1}} \cdots B_{\xi_1} y \,|\, S_{m-1} \Big\}$$

$$= \sum_{j=1}^{p} \alpha_j P\Big\{ B_{\xi_m} \text{ does not cover } z_j \Big\}$$

$$\leq \sum_{j=1}^{p} \alpha_j (q-1)/q = (q-1)/q, \tag{3.12}$$

where we have used (3.11). Apply the induction hypothesis to $P\{S_{m-1}\}$ and use (3.12) to prove (3.10).

Consider the infinite product

$$\cdots B_{\xi_m} B_{\xi_{m-1}} \cdots B_{\xi_1} x \quad (x \neq 0).$$

Partition the product into blocks as follows: B_{ξ_1} is the first block, $B_{\xi_3} B_{\xi_2}$ is the second block, and so on, with the property that the mth block has length m for $m = 1, 2, \ldots$. Let $B_{\xi_u} B_{\xi_{u-1}} \cdots B_{\xi_v}$ be the mth block (here $v < u$ depend on m). Denote by $T_m = T_m(x)$ the event

$$P\Big\{ B_{\xi_v} \text{ does not cover } B_{\xi_{v-1}} \cdots B_1 x;\ B_{\xi_{v+1}} \text{ does not cover}$$

$$B_{\xi_v} B_{\xi_{v-1}} \cdots B_1 x; \ldots, B_{\xi_u} \text{ does not cover } B_{\xi_{u-1}} B_{\xi_{u-2}} \cdots B_1 x \Big\}. \tag{3.13}$$

In particular, if T_m happens then the vectors $B_{\xi_{v-1}} \ldots B_1 x, \ldots, B_{\xi_u} B_{\xi_{u-1}} \ldots B_1 x$ are nonzero. By (3.10),

$$P\{T_m\} \leq ((q-1)/q)^m.$$

Since $\sum_{m=1}^{\infty} ((q-1)/q)^m < \infty$, by the Borel–Cantelli lemma with probability 1 only a finite number of the events $T_1, T_2, \ldots$ happen. Thus, with probability 1 there are infinitely many blocks (say, the blocks numbered $m_1 < m_2 < m_3 < \cdots$) such that T_{m_j} does not happen. But T_{m_j} does not happen precisely when either $B_{\xi_{v-1}} \cdots B_1 x = 0$ or at least one of B_{ξ_j} $(j = v, v+1, \ldots, u)$ covers $B_{\xi_{j-1}} \cdots B_1 x$ (we denote here the m_jth block by $B_{\xi_u} B_{\xi_{u-1}} \cdots B_{\xi_v}$, where $v < u$ depend on m_j). In the latter case we have

$$\|B_{\xi_j} B_{\xi_{j-1}} \cdots B_1 x\| \leq \alpha \|B_{\xi_{j-1}} \cdots B_1 x\|.$$

So, if Q_{m_j} stands for the product of the first m_j blocks, we have

$$\|Q_{m_j} x\| \leq \alpha^j \|x\|$$

(here we have used the inequalities $\|B_j\| \leq 1$). This proves that (3.9) happens with probability 1.

Apply (3.9) to a dense countable set D on the unit sphere. Using the countable additivity of the probability measure, it follows that there exists an infinite word w such that

$$\lim_{k \to \infty} B_{w^{(k)}} x = 0 \ \text{ for every } x \in D. \tag{3.14}$$

(In fact, we have proved more; namely, (3.14) holds with probability 1 when w is considered a random infinite word.) Since $\|B_j\| \leq 1$ for $j = 1, \ldots, q$, it follows that (3.14) holds for all x on the unit sphere. This proves Lemma 3.3. □

Proof of Theorem 1.1. In view of Lemmas 3.1, 3.2, and 3.3, we only need to verify that if $\mathcal{A} = \{A_i : i \in K\}$ is pointwise convergent and has the form (3.1), then for $j = 1, \dots, m$, the diagonal block $\mathcal{A}_j = \{A_{jj}^{(i)} : i \in K\}$ is a pointwise convergent set. But this statement is obvious from the definition of pointwise convergence. □

4. Some generalizations and extensions. The methods employed in this paper, especially the proof of Lemma 3.3, can be applied to obtain analogous results in other frameworks involving the notions of pointwise and uniform convergence. In this section we present several such results.

4.1. Nonlinear maps in F^n. Let $\mathcal{A} = \{A_i : i \in K\}$ be a set of nonlinear maps $A_i : F^n \to F^n$ (as before, $F = \mathbf{R}$ or $F = \mathbf{C}$). The definitions of uniform convergence and pointwise convergence given in section 1 carry over to this situation, replacing the product of matrices by superposition of nonlinear maps.

A map $A : F^n \to F^n$ is called *homogeneous* if $A(cx) = cA(x)$ for all $x \in F^n$ and $c \in F$. Clearly, a composition of homogeneous continuous maps is again homogeneous and continuous. Also, a homogeneous continuous map $A : F^n \to F^n$ is *Lipschitz continuous*; i.e., there is a constant $C > 0$ such that

$$\|A(x) - A(y)\| \le C\|x - y\|$$

for all $x, y \in F^n$; here $\|\cdot\|$ is some norm in F^n. It is easy to see that the notion of Lipschitz continuity is independent of the choice of the norm.

THEOREM 4.1. *Let $\mathcal{A} = \{A_i : i \in K\}$ be a set of homogeneous continuous maps $A_i : F^n \to F^n$. Assume that $\mathcal{A}$ generates a bounded semigroup (with superposition as the algebraic operation in the semigroup). Then $\mathcal{A}$ is pointwise convergent if and only if $\mathcal{A}$ is uniformly convergent.*

The proof of Theorem 4.1 is essentially the same as that of Lemma 3.3.

4.2. An abstract formulation. Let X be a metric space with the metric $d(x, y)$. Consider a set $\mathcal{A} = \{A_i : i \in K\}$ of continuous functions $X \to X$. We say that $\mathcal{A}$ is *pointwise convergent* to $x_0 \in X$ if for every $x \in X$ there is an infinite word w (which depends on x) such that

$$\lim_{k\to\infty} d\left(A_{w(k)}\left(A_{w(k-1)} \cdots \left(A_{w(1)}x\right)\cdots\right), x_0\right) = 0. \tag{4.1}$$

We say that $\mathcal{A}$ is *uniformly convergent* to x_0 if there is an infinite word w such that (4.1) holds for all $x \in X$.

THEOREM 4.2. *Let X be a compact separable metric space, and let $\mathcal{A}$ be a set of continuous functions $X \to X$. Assume, in addition, that*

$$d(A_i x, x_0) \le d(x, x_0)$$

for all $x \in X$ and all $A_i \in \mathcal{A}$. (In particular, x_0 is a fixed point for every $A_i \in \mathcal{A}$.) Then $\mathcal{A}$ is pointwise convergent to x_0 if and only if $\mathcal{A}$ is uniformly convergent to x_0.

The proof again follows by repeating the arguments used in the proof of Lemma 3.3.

REFERENCES

[G] L. N. GURVITS, *Stability of discrete linear inclusions*, Linear Algebra Appl., 231 (1995), pp. 47–85.

[DL1] I. Daubechies and J. C. Lagarias, *Sets of matrices all infinite products of which converge*, Linear Algebra Appl., 161 (1992), pp. 227–263.

[DL2] I. Daubechies and J. C. Lagarias, *Two scale difference equations,* II. *Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.

[BW] M. A. Berger and Y. Wang, *Bounded semigroups of matrices*, Linear Algebra Appl., 166 (1992), pp. 21–27.

[S] D. P. Stanford, *Stability for a multi-rate sampled-data system*, SIAM J. Control Optim., 17 (1979), pp. 390–399.

[SU] D. P. Stanford and J. M. Urbano, *Some convergence properties of matrix sets*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1132–1140.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, pp. 369–392, April 1997

009

COMPLETE ORTHOGONAL DECOMPOSITION FOR WEIGHTED LEAST SQUARES*

PATRICIA D. HOUGH† AND STEPHEN A. VAVASIS‡

Abstract. This paper proposes a complete orthogonal decomposition (COD) algorithm for solving weighted least-squares problems. In applications, the weight matrix can be highly ill conditioned, and this can cause standard methods like QR factorization to return inaccurate answers in floating-point arithmetic. Stewart and Todd independently established a norm bound for the weighted least-squares problem that is independent of the weight matrix. Vavasis proposed a definition of a "stable" solution of weighted least squares based on this norm bound: The solution computed by a stable algorithm must satisfy an accuracy bound that is not affected by ill conditioning in the weight matrix. A forward error analysis shows that the COD algorithm is stable in this sense, but it is simpler and more efficient than the algorithm proposed by Vavasis. Our forward error bound is contrasted to the backward error analysis of other previous works on weighted least squares.

Key words. weighted least squares, equilibrium systems, QR factorization, numerical stability, forward error analysis, interior-point methods

AMS subject classifications. 65F05, 65F25, 65G05

PII. S089547989528079X

1. Introduction.

We consider solving the problem

$$\min_{\mathbf{y}\in\mathbb{R}^n} \|D^{-1/2}(A\mathbf{y}-\mathbf{b})\| \tag{1.1}$$

for $\mathbf{y}$, where $D \in \mathbb{R}^{m\times m}$ is symmetric and positive definite, $A \in \mathbb{R}^{m\times n}$, $\mathbf{y} \in \mathbb{R}^n$, and $\mathbf{b} \in \mathbb{R}^m$. Two equivalent ways to write this problem are

$$A^T D^{-1} A\mathbf{y} = A^T D^{-1}\mathbf{b}$$

and

$$\begin{bmatrix} D & -A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix},$$

which is a special case of an equilibrium system. Applications of equilibrium systems include optimization, finite elements, structural analysis, and electrical networks [24]. These applications are discussed in more detail in section 2.

The following assumptions are made throughout the paper.

A1. *A has rank n, i.e., full column rank.*

A2. *D is diagonal.*

A1 and A2 imply that (1.1) is a full-rank weighted least-squares problem with a unique solution, and they allow the use of the norm bound obtained independently by Stewart [23] and Todd [25]. It should be noted that a similar result appears in a

*Received by the editors January 27, 1995; accepted for publication (in revised form) by D. P. O'Leary May 17, 1996. This work was supported by an NSF Presidential Young Investigator grant, with matching funds received from AT&T and Xerox Corporation. Additional support was received from ONR grant N00014-96-1-0050 and NSF grant DMS-9505155.

http://www.siam.org/journals/simax/18-2/28079.html

†Center for Applied Mathematics, Cornell University, Ithaca, NY 14853 (ph@cam.cornell.edu).

‡Department of Computer Science, Cornell University, Ithaca, NY 14853 (vavasis@cs.cornell.edu).

number of other works. See Forsgren [5] for a list of references. That bound is given in the following theorem.

THEOREM 1.1 (see [23], [25]). *Let $\mathcal{D}$ denote the set of all positive-definite $m \times m$ real diagonal matrices. Let A be an $m \times n$ real matrix of rank n. If we define χ_A and $\bar{\chi}_A$ as follows,*

$$\text{(a) } \chi_A = \sup\{\| \left(A^T D^{-1} A\right)^{-1} A^T D^{-1} \| : D \in \mathcal{D}\} \text{ and}$$
$$\text{(b) } \bar{\chi}_A = \sup\{\|A \left(A^T D^{-1} A\right)^{-1} A^T D^{-1} \| : D \in \mathcal{D}\},$$

then both χ_A and $\bar{\chi}_A$ are finite.

In this theorem, the norm can be any matrix norm induced by a vector norm. However, in this paper, $\|\cdot\| = \|\cdot\|_2$. Similarly, the condition number of a matrix M is the condition number of M in the 2-norm; i.e., $\kappa(M) = \kappa_2(M)$. We make one more assumption.

A3. *D is very ill conditioned.*

A discussion of the ill conditioning of D in applications is included in section 2. This assumption indicates that the coefficient matrix of the least-squares problem can also be ill conditioned. For this reason, the methods typically used to solve least-squares problems can give highly inaccurate solutions $\mathbf{y}$, as argued by Vavasis [28]. Since D is ill conditioned, we use the following definition of stability.

DEFINITION 1.2 (see [28]). *An algorithm for* (1.1) *is* stable *if, in the presence of finite-precision arithmetic, an error bound of the form*

$$\|\mathbf{y} - \hat{\mathbf{y}}\| \leq \epsilon \cdot f(A) \cdot \|\mathbf{b}\| \tag{1.2}$$

is satisfied, where $\mathbf{y}$ is the true solution, $\hat{\mathbf{y}}$ is the computed solution, $f(A)$ is some function of A not depending on D, and $\epsilon > 0$ is machine roundoff.

For the purposes of the upcoming analysis, other standard terminology is modified analogously. For example, a well-conditioned matrix is one for which there is an upper bound on the condition number that does not depend on D. In order to show that the proposed algorithm is stable, then, we strive to obtain bounds on norms, condition numbers, and errors that do not depend on D.

We now present the algorithm.

ALGORITHM: Complete orthogonal decomposition (COD).

Step 1: QR factor, with column pivoting, $A^T D^{-1/2}$ to get

$$A^T D^{-1/2} = QRP, \tag{1.3}$$

where Q is an $n \times n$ orthogonal matrix, R is an $n \times m$ upper triangular ("trapezoidal") matrix, and P is an $m \times m$ permutation matrix.

Step 2: Apply reduced QR factorization (without pivoting) to R^T to get

$$R^T = Z_1 U_1, \tag{1.4}$$

where Z_1 is an $m \times n$ matrix with orthonormal columns and U_1 is an $n \times n$ upper triangular matrix.

Step 3: Solve the following system, via backsubstitution, for $\bar{\mathbf{y}}$:

$$U_1 \bar{\mathbf{y}} = Z_1^T P D^{-1/2} \mathbf{b}. \tag{1.5}$$

Step 4: To get $\mathbf{y}$, multiply the result of Step 3 by Q:

$$\mathbf{y} = Q\bar{\mathbf{y}}. \tag{1.6}$$

Both QR factorizations can be computed with Householder reflections, although other methods are also acceptable. Note that the QR factorization for the least-squares problem occurs in Step 2. The QR factorization in Step 1 is to make the algorithm stable. The solution of least-squares problems via QR factorization with column pivoting was introduced by Golub [9]. The term "complete orthogonal decomposition" refers to a factorization of the form QRZ in which Q and Z are orthogonal and R is triangular [10]. Therefore, we have chosen this name for the above algorithm, which computes a particular kind of COD.

In exact arithmetic, the COD algorithm solves the weighted least-squares problem given by (1.1). Writing the problem as a system of equations gives

$$D^{-1/2}A\mathbf{y} \overset{LS}{=} D^{-1/2}\mathbf{b}.$$

After performing the QR factorization in Step 1, $D^{-1/2}A$ can be replaced by the equivalent quantity $P^TR^TQ^T$. The system of equations becomes

$$P^TR^TQ^T\mathbf{y} \overset{LS}{=} D^{-1/2}\mathbf{b}$$

or, equivalently,

$$R^TQ^T\mathbf{y} \overset{LS}{=} PD^{-1/2}\mathbf{b}.$$

Letting $\bar{\mathbf{y}} = Q^T\mathbf{y}$ constitutes a change of variables and transforms the above system of equations into

$$R^T\bar{\mathbf{y}} \overset{LS}{=} PD^{-1/2}\mathbf{b},$$

which is again a least-squares problem. Steps 2 and 3 are a standard method for solving least-squares problems, so the result in exact arithmetic is the solution $\bar{\mathbf{y}}$ to the transformed problem.

The most common methods for solving weighted least-squares problems are the same as those used to solve unweighted problems. These include QR factorization and solving the normal equations via Cholesky factorization. There are also some specialized algorithms for weighted least squares that will be discussed in section 7.

Most of this paper is devoted to an analysis of the stability of the COD algorithm. Before giving a rigorous stability analysis of the algorithm, we offer an intuitive explanation of why this algorithm finds an accurate solution. The first step is a QR factorization of a matrix that is well conditioned up to a scaling of the columns. So the result is a computed upper triangular matrix that is close to the exact upper triangular matrix. It would be useful to know something about the condition number of this matrix as well. To minimize confusion assume, without loss of generality, that $A^TD^{-1/2}$ has been "prepivoted." This means that the columns of $A^TD^{-1/2}$ are ordered in such a way that the norms of the first n columns are, loosely speaking, in decreasing order. In addition, the norms of the first n columns are larger than those of the last $m-n$ columns. One might suspect that this implies that the entries of $D^{-1/2}$ are ordered in the same way. In other words, some inequality similar to

$$d_i^{-1/2} \geq d_j^{-1/2} \quad \text{for } i \leq j, 1 \leq i \leq n, 1 \leq j \leq m$$

might hold. This ordering becomes significant in the second step of the algorithm.

Recall that

$$R = Q^T A^T D^{-1/2}.$$

Notice that $Q^T A^T$ is upper triangular. So let

$$\bar{R} = Q^T A^T.$$

Notice also that R is ill conditioned and that the ill conditioning arises from $D^{-1/2}$. We try to "offset" the effects of $D^{-1/2}$ in the following naive way. Let $\bar{D} = D(1:n, 1:n)$ and consider the following:

$$\bar{D}^{1/2} R = \bar{D}^{1/2} \bar{R} D^{-1/2} = \begin{bmatrix} \left(\frac{d_1}{d_1}\right)^{1/2} \bar{r}_{11} & \cdots & \left(\frac{d_1}{d_n}\right)^{1/2} \bar{r}_{1n} & \cdots & \left(\frac{d_1}{d_m}\right)^{1/2} \bar{r}_{1m} \\ & \ddots & \vdots & & \vdots \\ & & \left(\frac{d_n}{d_n}\right)^{1/2} \bar{r}_{nn} & \cdots & \left(\frac{d_n}{d_m}\right)^{1/2} \bar{r}_{nm} \end{bmatrix}.$$

Recall that in this paper "well conditioned" means that the matrix under consideration has a condition number determined by a property of A independently of D, because our focus is on the ill conditioning in D. Thus, $\bar{R}$ is trivially well conditioned since it has the same singular values as A. If the weights are indeed in the order described above, then it is not difficult to show that there are upper bounds on all entries of $\bar{D}^{1/2} R$. It can also be shown that there are upper bounds on the entries of $\left(\bar{D}^{1/2} R(:, 1:n)\right)^{-1}$. Using this information, it is not difficult to show that $\bar{D}^{1/2} R$ (and hence $R^T \bar{D}^{1/2}$) is well conditioned; i.e., R^T is well conditioned up to a scaling of the columns. In the second step, then, we have a least-squares problem with a coefficient matrix that is well conditioned up to a scaling of the columns; namely, solve

$$\min_{\bar{\mathbf{y}} \in \mathbb{R}^n} \|R^T \bar{\mathbf{y}} - D^{-1/2} \mathbf{b}\|$$

for $\bar{\mathbf{y}}$. This yields an upper triangular matrix U_1 in (1.4) that is also well conditioned up to scaling of the columns. We show that U_1 is also well conditioned up to scaling of the rows (see (5.1) below). In traditional analysis, U_1 being well conditioned up to a scaling of rows indicates that the standard algorithms for solving (1.5) give an accurate solution (e.g., inequality (3.1.1) from [10] combined with Theorem 2.7.3 of [10] shows that scaling rows does not change the error bounds).

The remainder of the paper contains a detailed discussion of the topics mentioned in this section. Applications of weighted least-squares problems are described in section 2, and the relevance of a forward error bound in terms of each specific application is explained. Then a discussion in section 3 of a numerical issue, namely, checking for linear dependence among the rows of A, leads into the rigorous forward error analysis of the COD algorithm in sections 4 through 6. In section 7, the analysis is then compared to the (backward error) analyses of algorithms found in the literature. The paper concludes with a discussion of open questions.

2. Applications of stable weighted least squares. In this section we discuss three applications of (1.1) that involve ill-conditioned weight matrices D and explain the role of our forward error bound (1.2) in these applications. In two of the three

applications, we provide summaries of computational experiments from our previous work.

Electrical networks. Perhaps the simplest application of (1.1) is to resistive networks with fixed voltage sources (batteries). In this case, D encodes the resistances of the resistors in the network, $\mathbf{y}$ encodes the voltages of the nodes, A encodes node-wire adjacency (A is a node-arc adjacency matrix; all of its entries are either 0, 1, or -1), and $\mathbf{b}$ encodes battery voltages.

The case when D is ill conditioned arises when there are highly varying resistances in the circuit. Highly varying resistances can occur, for instance, when one tries to model current leakage through insulators as part of the problem.

We have conducted small-scale computational experiments on electrical networks using the COD algorithm and the nullspace-scaled hybrid (NSH) method [28], which also satisfies (1.2). For node-arc adjacency matrices, χ_A and $\bar{\chi}_A$ are bounded by $O(m)$, so (1.2) states that the voltages should be calculated to within machine precision multiplied by the norm of the battery voltages. When applied to such problems, the NSH and the COD algorithms yield solutions with 15 digits of accuracy, while textbook methods (e.g., [3]) routinely give answers without any significant digits of accuracy.

Finite element methods. A more sophisticated application of (1.1) is to solve boundary value problems of the form

$$\begin{aligned} \nabla \cdot (c \nabla u) &= 0 \quad \text{on } \Omega, \\ u &= g \quad \text{on } \partial\Omega \end{aligned}$$

for u. Here, c is a user-specified conductivity field on the domain Ω that must be positive at every point. Function g is the user-specified Dirichlet boundary data. This boundary value problem arises in many fields of science and engineering; an example application is the heat equilibrium equation, in which case u stands for the steady-state temperature field in the domain and c stands for thermal conductivity.

As argued in [27], it is possible to express the standard piecewise-linear finite element method for this boundary value problem as a weighted least-squares problem. The diagonal weight matrix D encodes the conductivity field c. Matrix A encodes geometric information about the triangulation. Vector $\mathbf{y}$ stands for the solution field u, and vector $\mathbf{b}$ encodes the Dirichlet boundary data. Thus, the case when D is ill conditioned corresponds to finite element problems with highly varying conductivity. This in turn corresponds to applications in which the domain is composed of varying materials.

The analysis in [27] shows that variants of $\chi_A, \bar{\chi}_A$ are modest for this geometry matrix A, at least in the case of finite element triangulations with bounded aspect ratio and with all dihedral angles bounded by $\pi/2$. Thus, the stability bound (1.2) states that our method should compute the finite element solution $\mathbf{y}$ to all significant digits of accuracy, relative to the boundary data. This solution $\mathbf{y}$ is still inexact for u because of truncation error that is always present in finite element methods; our analysis addresses roundoff error rather than truncation error.

The computational experiments presented in [27] are larger-scale computations. In a problem with conductivity varying by 15 orders of magnitude, the traditional finite element solution method returns an answer with no significant digits because of roundoff error. In contrast, a variant of the NSH method (which takes advantage of the isotropic nature of the problem) [27] that satisfies our stability bound (1.2) returns an answer which was as accurate as could be hoped for compared to an analytically-

derived solution, given the presence of truncation error. The COD method of this paper would require a similar modification to take advantage of isotropy.

Interior-point methods. It is well known that the system of equations for the Newton step in an interior-point method can be expressed as a weighted least-squares problem [11] of the form (1.1). To be precise, consider the linear programming problem

$$\begin{array}{rl} \text{minimize} & \mathbf{c}^T\mathbf{x} \\ \text{subject to} & A^T\mathbf{x} = \mathbf{b}, \\ & \mathbf{x} \geq \mathbf{0} \end{array}$$

whose dual is

$$\begin{array}{rl} \text{maximize} & \mathbf{b}^T\mathbf{y} \\ \text{subject to} & A\mathbf{y} + \mathbf{s} = \mathbf{c}, \\ & \mathbf{s} \geq \mathbf{0} \end{array}$$

(which is standard form, except that we have transposed A to be consistent with least-squares notation). A standard primal-dual method starting at a feasible interior point $(\mathbf{x}, \mathbf{y}, \mathbf{s})$ for this problem computes an update to $\mathbf{y}$ of the form

$$A^T D A \Delta\mathbf{y} = A^T D(\mathbf{s} - \sigma\mu X^{-1}\mathbf{e}), \tag{2.1}$$

where $X = \operatorname{diag}(\mathbf{x})$, $S = \operatorname{diag}(\mathbf{s})$, $D = S^{-1}X$, σ is an algorithm-dependent parameter, μ is the duality gap, and $\mathbf{e}$ is the vector of all ones [31]. Thus, (2.1) is the system of normal equations for a weighted least-squares problem to compute $\Delta\mathbf{y}$, and therefore our accuracy bound (1.2) implies that algorithm COD computes $\Delta\mathbf{y}$ accurately with respect to $\mathbf{s}$. (The other term $\mu X^{-1}\mathbf{e}$ in (2.1) is on the same order as $\mathbf{s}$ because of proximity to the central path.)

Interior-point methods are an especially interesting application of weighted least squares for several reasons. First, the distance to some of the constraints must tend to zero. This means that ill conditioning in D *always* occurs, unlike in the preceding application domains. Another interesting difference is that in the interior-point method we need not only $\Delta\mathbf{y}$, but also updates to $\mathbf{x}$ and $\mathbf{s}$ usually denoted as $\Delta\mathbf{x}$ and $\Delta\mathbf{s}$. These are obtained from $\Delta\mathbf{y}$; for example, $\Delta\mathbf{s} = -A\Delta\mathbf{y}$. In fact, our forward error bound is not sufficiently strong to obtain the requisite accuracy bound on $\Delta\mathbf{x}$ or $\Delta\mathbf{s}$ because some components are very small as convergence is achieved [30]. This means that there is a demand for more accuracy in some components of $A\Delta\mathbf{y}$ than what could be obtained from our forward error bound (1.2). A final issue with interior-point methods is that, unlike the two preceding applications, we have no estimates a priori of χ_A or $\bar{\chi}_A$. The COD algorithm does not require estimates of these parameters, but without knowledge of the parameters we cannot evaluate the strength of our stability bound (1.2). For certain special classes of linear programming problems such as network optimization, prior estimates are possible.

We return to the topic of interior-point methods in section 8.

3. A note on numerical tolerance. In the upcoming analysis we assume throughout that any occurrence of exact linear dependence among the columns of A^T is always determined correctly in Step 1 of the algorithm (QR factorization with pivoting). This requires the use of a numerical tolerance. To illustrate this point, consider applying the algorithm when $D^{-1/2} = \operatorname{diag}(1, 1, 1, 10^{-20})$ and

$$A^T = \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & 1 & -1 & 0 \\ 1 & 0 & 1 & 7 \end{pmatrix}.$$

Observe that the third column of A^T is dependent on the first two. If the QR factorization of $A^T D^{-1/2}$ were done in exact arithmetic, this dependence would be manifested as a "0" in the $(3,3)$ position of the factored $A^T D^{-1/2}$ after the first two QR factorization steps, and column 4 would be chosen for the third pivot.

In finite-precision arithmetic, however, we would expect the $(3,3)$ entry to be on the order of machine-epsilon rather than 0. Because column 4 is weighted by 10^{-20}, the unwanted residual in the $(3,3)$ position could cause column 3 to be chosen for the third column pivot instead of column 4. Thus, without modification, ordinary QR factorization with column pivoting procedure has missed a linear dependence.

We address this problem as follows: after the kth QR factorization step, we check whether the residual portion (that is, positions $k+1,\ldots,n$) of any uneliminated column has become very small (according to some tolerance level) with respect to the original norm of that column. If so, those entries are changed to zeros. Notice that this test requires very little additional work because the usual QR factorization algorithm with column pivoting already monitors the norms of the residual portions of the columns [9]. In this way, if there are exact dependences among the rows of A, the algorithm does not miss them.

If this numerical test fails to detect exact dependence, then the following stability analysis no longer holds. It can be shown that the test we have proposed will fail only in the case that there is near-dependence among the columns of A^T. However, in this case, the parameter χ_A given by Theorem 1.1 is large, and so the stability bound (which depends on χ_A and $\bar{\chi}_A$) is not practically applicable.

4. The first QR factorization. The intuitive discussion in section 1 asserted that the ordering of the weights produced in Step 1 of the algorithm is important in stabilizing the algorithm. It is necessary, then, to establish this order. The pivoting in Step 1 chooses a particular set of rows of $D^{-1/2}A$. The corresponding rows of A form a basis for the row space of A. Thus, we will refer to the rows chosen by the column pivoting as the "basis rows" of A. We must determine how the weight of a particular basis row compares to those of other rows in the same basis and to those of the rows not in that basis. To do this, we start with a general result about any set of rows that forms a basis for the row space of A.

LEMMA 4.1. *Let B be an $n \times n$ matrix whose columns are an arbitrary set of n rows $\mathbf{a}_{i_1}^T,\ldots,\mathbf{a}_{i_n}^T$ of A that form a basis for the row space of A. Then*

$$\max_{1\le j\le n} \|\mathbf{a}_{i_j}\| \le (\chi_A \|A\|) \cdot \min_{1\le j\le n} \|\mathbf{a}_{i_j}\|. \tag{4.1}$$

Proof. Let B and $\mathbf{a}_{i_1}^T,\ldots,\mathbf{a}_{i_n}^T$ be as in the lemma. Without loss of generality, suppose that $\|\mathbf{a}_{i_n}\| = \min_{1\le j\le n} \|\mathbf{a}_{i_j}\|$. Then write

$$B = \left[\hat{B}, \mathbf{a}_{i_n}\right]$$

and partition B^{-1} as

$$B^{-1} = \begin{bmatrix} X \\ \mathbf{v}^T \end{bmatrix}.$$

Then

$$B^{-1}B = I = \begin{bmatrix} X\hat{B} & X\mathbf{a}_{i_n} \\ \mathbf{v}^T\hat{B} & \mathbf{v}^T\mathbf{a}_{i_n} \end{bmatrix}.$$

This means $\mathbf{v}^T\mathbf{a}_{i_n} = 1$. By the Cauchy–Schwarz inequality, $\|\mathbf{v}\| \geq 1/\|\mathbf{a}_{i_n}\|$. Also, $\|\mathbf{v}\| \leq \|B^{-1}\|$ because $\mathbf{v}$ is a row of B^{-1}, and

$$\|B^{-1}\| \leq \chi_A \tag{4.2}$$

(see [28] for the proof of (4.2)). Combining these inequalities yields

$$1/\|\mathbf{a}_{i_n}\| \leq \chi_A;$$

i.e.,

$$\min_{1\leq j\leq n} \|\mathbf{a}_{i_j}\| \geq 1/\chi_A.$$

Multiply both sides of this inequality by the inequality $\|A\| \geq \max_{1\leq j\leq n} \|\mathbf{a}_{i_j}\|$ to obtain

$$\max_{1\leq j\leq n} \|\mathbf{a}_{i_j}\| \leq (\chi_A\|A\|) \cdot \min_{1\leq j\leq n} \|\mathbf{a}_{i_j}\|,$$

as required. □

Notice that the above lemma implies that there is a lower bound on the norm of every column of any basis for the row space of A.

Suppose now that $k > 0$ steps of the factorization have been completed. Partition the resulting matrix by rows as follows:

$$AQ_1 \cdots Q_k = \bar{A} = \begin{bmatrix} \alpha_{11} & 0 & \cdots & \mathbf{0}^T \\ \vdots & \ddots & \ddots & \vdots \\ \alpha_{k1} & \cdots & \alpha_{kk} & \mathbf{0}^T \\ \alpha_{k+1,1} & \cdots & \alpha_{k+1,k} & \bar{\mathbf{a}}_{k+1}^T \\ \vdots & & \vdots & \vdots \\ \alpha_{m,1} & \cdots & \alpha_{m,k} & \bar{\mathbf{a}}_m^T \end{bmatrix}.$$

Lemma 4.1 can be extended so that it applies to the residual portions of the rows of $\bar{A}$, i.e., $\bar{\mathbf{a}}_{k+1}^T, \ldots, \bar{\mathbf{a}}_n^T$, as follows.

LEMMA 4.2. *Let B be an $n \times n$ matrix whose first k columns are the k rows of A, say $\mathbf{a}_{i_1}^T, \ldots, \mathbf{a}_{i_k}^T$, chosen by the column pivoting in the first k steps of the QR factorization. Let the remaining columns of B be arbitrary rows of A, say $\mathbf{a}_{i_{k+1}}^T, \ldots, \mathbf{a}_{i_n}^T$, such that the columns of B form a basis for the row space of A. As with $\bar{A}$ above, write*

$$Q_k^T \cdots Q_1^T B = \begin{bmatrix} \alpha_{i_1,1} & \cdots & \alpha_{i_k,1} & \alpha_{i_{k+1},1} & \cdots & \alpha_{i_n,1} \\ 0 & \ddots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & \alpha_{i_k k} & \alpha_{i_{k+1} k} & & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \bar{\mathbf{a}}_{i_{k+1}} & \cdots & \bar{\mathbf{a}}_{i_n} \end{bmatrix}.$$

Then

$$\max_{k+1\leq j\leq n} \|\bar{\mathbf{a}}_{i_j}\| \leq (\chi_A\|A\|) \cdot \min_{k+1\leq j\leq n} \|\bar{\mathbf{a}}_{i_j}\|. \tag{4.3}$$

Proof. Let $Q = Q_1 \cdots Q_k$, and let B be defined as above. Then

$$Q^T B = \begin{bmatrix} R & X \\ 0 & \bar{B} \end{bmatrix},$$

where R is upper triangular. Comparing this matrix to the partitioned one above, we see that the columns of $\bar{B}$ are $\bar{\mathbf{a}}_{i_{k+1}}, \ldots, \bar{\mathbf{a}}_{i_n}$. To prove Lemma 4.2, then, we need a lower bound on a typical column of $\bar{B}$. Notice that

$$(Q^T B)^{-1} = \begin{bmatrix} R^{-1} & -R^{-1} X \bar{B}^{-1} \\ 0 & \bar{B}^{-1} \end{bmatrix}.$$

Therefore,

$$\|\bar{B}^{-1}\| \leq \|B^{-1}\| \leq \chi_A.$$

Now, as above, turn an upper bound on $\|\bar{B}^{-1}\|$ into a lower bound on any column of $\bar{B}$. □

Using the previous two lemmas, we can determine the relationships between the weights of the basis rows and those of the nonbasis rows. For the remainder of the paper assume, as in the intuitive discussion in section 1, that the columns of $A^T D^{-1/2}$ have been reordered so that no pivoting is necessary. This implies that not only do the first n columns of A^T form a basis for the column space of A^T, but that there is also an order that has been imposed on the columns of A^T. Notice that the "prepivoting" also implies that $\bar{\mathbf{a}}_{i_j}$ becomes $\bar{\mathbf{a}}_j$ for the remainder of the paper. So let B be the $n \times n$ matrix whose columns are rows $\mathbf{a}_1^T, \ldots, \mathbf{a}_n^T$ and let $d_1, \ldots, d_m$ denote the (reordered) entries of D.

THEOREM 4.3. *Suppose the first $k \geq 0$ steps of the QR factorization have been completed. If $d_{k+1}^{-1/2}$ is the weight assigned to $\mathbf{a}_{k+1} \in B$ and $d_j^{-1/2}$ is the weight assigned to $\mathbf{a}_j \notin B$, then*

$$\frac{d_{k+1}}{d_j} \leq (\chi_A \|A\|)^4, \tag{4.4}$$

provided $\mathbf{a}_j$ is linearly independent of the first k basis vectors.

Proof. For $k \geq 0$, let $\bar{A}$ be as defined before Lemma 4.2. (Notice that when $k = 0$ this is just the matrix A partitioned into rows.) Since the columns of B form a basis for the row space of A,

$$\mathbf{a}_j = \sum_{i=1}^{n} c_i \mathbf{a}_i.$$

Assuming $\mathbf{a}_j$ is linearly independent of the first k basis rows implies

$$\mathbf{a}_j - \sum_{i=1}^{k} c_i \mathbf{a}_i = \sum_{i=k+1}^{n} c_i \mathbf{a}_i \neq \mathbf{0},$$

which means that $c_i \neq 0$ for at least one i such that $k+1 \leq i \leq n$. Take $Q = Q_1 \cdots Q_k$. Then

$$Q^T \mathbf{a}_j - \sum_{i=1}^{k} c_i Q^T \mathbf{a}_i = \sum_{i=k+1}^{n} c_i Q^T \mathbf{a}_i$$

and

$$\bar{\mathbf{a}}_j - \sum_{i=1}^{k} c_i \bar{\mathbf{a}}_i = \sum_{i=k+1}^{n} c_i \bar{\mathbf{a}}_i,$$

where $\bar{\mathbf{a}}_i$ is the residual portion of $\mathbf{a}_i$. Notice that $\bar{\mathbf{a}}_i = \mathbf{0}$ for $1 \le i \le k$. So

$$\bar{\mathbf{a}}_j = \sum_{i=k+1}^{n} c_i \bar{\mathbf{a}}_i,$$

where $c_i \neq 0$ for at least one i. Let l be such that $k+1 \le l \le n$ and $c_l \neq 0$. Then

$$\bar{\mathbf{a}}_l = \frac{1}{c_l} \left(\bar{\mathbf{a}}_j - \sum_{i=k+1, i \neq l}^{n} c_i \bar{\mathbf{a}}_i \right).$$

So $\bar{B} = \{\bar{\mathbf{a}}_j, \bar{\mathbf{a}}_{k+1}, \ldots, \bar{\mathbf{a}}_{l-1}, \bar{\mathbf{a}}_{l+1}, \ldots, \bar{\mathbf{a}}_n\}$ is a basis for $\{\bar{\mathbf{a}}_{k+1}, \ldots, \bar{\mathbf{a}}_m\}$. Since (4.1) and (4.3) hold for any basis for the row space of A,

$$\|\bar{\mathbf{a}}_j\| \ge \frac{\max\left\{\|\bar{\mathbf{a}}_i\| : \bar{\mathbf{a}}_i \in \bar{B}\right\}}{\chi_A \|A\|}.$$

Recall that the columns of $A^T D^{-1/2}$ have been reordered so that no pivoting is necessary. This means that there is an order imposed on the columns of $A^T D^{-1/2}$. More specifically, at step $k+1$

$$\left(\frac{1}{d_j}\right)^{1/2} \|\bar{\mathbf{a}}_j\| \le \left(\frac{1}{d_{k+1}}\right)^{1/2} \|\bar{\mathbf{a}}_{k+1}\|.$$

Thus,

$$\begin{aligned}
\frac{d_{k+1}}{d_j} &\le \left(\frac{\|\bar{\mathbf{a}}_{k+1}\|}{\|\bar{\mathbf{a}}_j\|}\right)^2 \\
&\le \left(\frac{\chi_A \cdot \|A\| \cdot \max_{k+1 \le i \le n} \|\bar{\mathbf{a}}_i\|}{\max\left\{\|\bar{\mathbf{a}}_i\| : \bar{\mathbf{a}}_i \in \bar{B}\right\}}\right)^2 \\
&\le \left(\frac{\chi_A \cdot \|A\| \cdot \max_{k+1 \le i \le n} \|\bar{\mathbf{a}}_i\|}{\min_{k+1 \le i \le n} \|\bar{\mathbf{a}}_i\|}\right)^2 \\
&\le \left(\chi_A \|A\|\right)^4,
\end{aligned}$$

which is (4.4). ∎

It is also necessary to know the relationships between the weights of the basis rows of A. Suppose $\mathbf{a}_i, \mathbf{a}_j \in B$, where $i < j$. It follows from (4.1), (4.3), and the implicit order indicated by the absence of column pivoting that

$$\frac{d_i}{d_j} \le \left(\chi_A \|A\|\right)^2 \le \left(\chi_A \|A\|\right)^4. \tag{4.5}$$

Recall that the intuitive argument given in section 1 relied on the weights being in the following order:

$$d_i^{-1/2} \ge d_j^{-1/2} \quad \text{for } i \le j, 1 \le i \le n, 1 \le j \le m.$$

Theorem 4.3 indicates, however, that they are not ordered in exactly this way. Instead, this ordering holds up to scaling by a constant; i.e.,

$$d_i^{-1/2} \ge \frac{d_j^{-1/2}}{\left(\chi_A \|A\|\right)^2} \quad \text{for } i \le j, 1 \le i \le n, 1 \le j \le m.$$

This bound is sufficient for the arguments that follow.

The second step of the algorithm performs a QR factorization on R^T. To analyze that step, then, it is necessary to know something about the condition of R. The relationships between the weights of the rows of A are used in the proof of the following theorem, which states that R is well conditioned up to a scaling of the rows or the columns. Recall that for an $m \times n$ matrix M of rank n, $\kappa(M)$ is the condition number (in the 2-norm) of M; i.e.,

$$\kappa(M) = \|M\| \cdot \| \left(M^T M\right)^{-1} M^T \|.$$

THEOREM 4.4. *Let $C = \bar{D}^a R D^{1/2-a}$, where $a \geq 0$ and $\bar{D} = D(1:n, 1:n)$. If $\tilde{C} = C(1:k, :)$, then*

$$\kappa(\tilde{C}) \leq n^4 \cdot (\chi_A \|A\|)^{16a+2} \tag{4.6}$$

for any $1 \leq k \leq n$.

Proof. First, we must find an upper bound on $\|\tilde{C}\|$. Since $\tilde{C}$ is a submatrix of C, $\|\tilde{C}\| \leq \|C\|$. Therefore, it is sufficient to show that there is an upper bound on $\|C\|$. Write C as follows:

$$C = \bar{D}^a R D^{1/2-a} = \bar{D}^a Q^T A^T D^{-1/2} D^{1/2-a} = \bar{D}^a \bar{R} D^{-a},$$

where $\bar{R} = Q^T A^T = R D^{1/2}$. If the entries of C are written explicitly,

$$C = \begin{bmatrix} \left(\frac{d_1}{d_1}\right)^a \bar{r}_{11} & \cdots & \left(\frac{d_1}{d_n}\right)^a \bar{r}_{1n} & \cdots & \left(\frac{d_1}{d_m}\right)^a \bar{r}_{1m} \\ & \ddots & \vdots & & \vdots \\ & & \left(\frac{d_n}{d_n}\right)^a \bar{r}_{nn} & \cdots & \left(\frac{d_n}{d_m}\right)^a \bar{r}_{nm} \end{bmatrix}.$$

Consider $\bar{R}_1 = \bar{R}(:, 1:n)$. Again, let B be the basis consisting of the first n columns of A^T. Then $\bar{R}_1 = Q^T B$. So $|\frac{1}{\bar{r}_{ii}}| \leq \|B^{-1}\| \leq \chi_A$ for $1 \leq i \leq n$. If $\bar{\mathbf{r}}_i^T$ is the ith row of $\bar{R}$, then $\|\bar{\mathbf{r}}_i^T\| \leq \|A\|$ for all $1 \leq i \leq n$. These facts, (4.4), and (4.5) imply the following:

$$\frac{1}{\chi_A} \leq |\bar{r}_{ii}| \leq \|A\|, 1 \leq i \leq n \text{ and} \tag{4.7}$$

$$\left\| d_i^a \bar{\mathbf{r}}_j^T D^{-a} \right\| \leq \|A\| \cdot (\chi_A \|A\|)^{4a}, 1 \leq i \leq j \leq n. \tag{4.8}$$

Recall that Theorem 4.3 (and thus (4.8)) holds only when $\mathbf{a}_j$ is linearly independent of the first $i-1$ basis vectors. We must now consider the case not covered by Theorem 4.3. Suppose that B and D are defined as before. For each nonbasis row $\mathbf{a}_j$ there is a $1 \leq k \leq n$ such that $\mathbf{a}_j$ is linearly independent of the first $k-1$ basis vectors, but is linearly dependent on the first k basis vectors. So

$$\mathbf{a}_j = \sum_{i=1}^{k} c_i \mathbf{a}_i,$$

where $c_k \neq 0$. Now suppose that k steps of the QR factorization have been completed. Then

$$Q_k^T \cdots Q_1^T \mathbf{a}_j = \sum_{i=1}^{k} c_i Q_k^T \cdots Q_1^T \mathbf{a}_i = \sum_{i=1}^{k} c_i \begin{bmatrix} \alpha_{1i} \\ \vdots \\ \alpha_{ii} \\ \mathbf{0} \end{bmatrix}.$$

So $\bar{\mathbf{a}}_j = \mathbf{0}$ (where $\bar{\mathbf{a}}_j$ is as in Lemma 4.2). After this point, transformations act only on $\bar{\mathbf{a}}_j$. This gives

$$Q^T \mathbf{a}_j = \sum_{i=1}^{k} c_i \begin{bmatrix} \alpha_{1i} \\ \vdots \\ \alpha_{ii} \\ \mathbf{0} \end{bmatrix},$$

telling us that $\bar{r}_{ij} = 0$ for $i > k$, so these entries do not contribute to the norm in (4.8). Thus, (4.8) holds even in the case where $\mathbf{a}_j$ is not linearly independent of the first $i-1$ basis vectors.

If $\mathbf{c}_i^T$ is the ith row of C, then

$$\begin{aligned} \|C\| &\le \sum_{i=1}^{n} \|\mathbf{c}_i^T\| \\ &\le n \cdot \max_{1 \le i \le n} \|\mathbf{c}_i^T\| \\ &= n \cdot \max_{1 \le i \le n} \|d_i^a \bar{\mathbf{r}}_i^T D^{-a}\| \\ &\le n \cdot \|A\| \cdot (\chi_A \|A\|)^{4a}. \end{aligned} \tag{4.9}$$

The third line follows from the definition of C and the fourth line follows from (4.8). The next step is to find an upper bound on $\|(\tilde{C}\tilde{C}^T)^{-1}\|$. Let $\tilde{C}_1 = \tilde{C}(:, 1:k)$. Notice that

$$\|(\tilde{C}\tilde{C}^T)^{-1}\| \le \|\tilde{C}_1^{-T}\|^2.$$

If $C_1 = C(:, 1:n)$, it is easy to show that $\tilde{C}_1^{-T}$ is a submatrix of C_1^{-T}. To obtain an upper bound on $\|C_1^{-T}\|$, we use the following fact, which will be proved after the current proof.

Fact. If $C_1 = C(1:n, 1:n)$, then

$$\|C_1^{-T}\| \le n \cdot \chi_A \cdot (\chi_A \|A\|)^{4a}.$$

So

$$\begin{aligned} \|(\tilde{C}\tilde{C}^T)^{-1}\| &= \|\tilde{C}_1^{-T}\|^2 \\ &\le \|C_1^{-T}\|^2 \\ &\le \left[n \cdot \chi_A \cdot (\chi_A \|A\|)^{4a} \right]^2, \end{aligned}$$

and the bound on the condition number is

$$\begin{aligned} \kappa(\tilde{C}) &= \|\tilde{C}\| \cdot \|\tilde{C}^T (\tilde{C}\tilde{C}^T)^{-1}\| \\ &\le \|\tilde{C}\|^2 \cdot \|(\tilde{C}\tilde{C}^T)^{-1}\| \\ &\le n^4 \cdot (\chi_A \|A\|)^{16a+2}. \end{aligned}$$

Thus, the theorem is proved. □

The above theorem implies that R is not only well conditioned up to a scaling of the rows, but is also well conditioned up to a scaling of either the rows or the columns. This result will be useful later in the analysis.

Recall that we must still prove the fact used in the above proof. We state it in the form of the following lemma and give the proof below.

LEMMA 4.5. *Let $C_1 = C(:, 1:n)$, where C is defined as in the previous theorem. Then*

$$\|C_1^{-T}\| \le n \cdot \chi_A \cdot (\chi_A \|A\|)^{4a}. \tag{4.10}$$

Proof. Recall that

$$C = \bar{D}^a \bar{R} D^{-a},$$

where $\bar{D} = D(1:n, 1:n)$ and $\bar{R} = Q^T A^T = RD^{1/2}$. So

$$C_1 = \bar{D}^a \bar{R}(:, 1:n) \bar{D}^{-a}.$$

Notice that

$$\bar{R}(:, 1:n) = Q^T B,$$

where B is the row space basis consisting of the first n rows of A. If $L = Q^T B^{-T}$, then

$$\begin{aligned} C_1^{-T} &= \bar{D}^{-a} L \bar{D}^a \\ &= \begin{bmatrix} \left(\frac{d_1}{d_1}\right)^a l_{11} & & & \\ \left(\frac{d_1}{d_2}\right)^a l_{21} & \left(\frac{d_2}{d_2}\right)^a l_{22} & & \\ \vdots & \vdots & \ddots & \\ \left(\frac{d_1}{d_n}\right)^a l_{n1} & \left(\frac{d_2}{d_n}\right)^a l_{n2} & \cdots & \left(\frac{d_n}{d_n}\right)^a l_{nn} \end{bmatrix}. \end{aligned}$$

If $\mathbf{c}_i^T$ is the ith row of C_1^{-T}, then

$$\begin{aligned} \|C_1^{-T}\| &\le \sum_{i=1}^{n} \|\mathbf{c}_i^T\| \\ &\le n \cdot \max_{1 \le i \le n} \|\mathbf{c}_i^T\| \\ &\le n \cdot (\chi_A \|A\|)^{4a} \cdot \max_{1 \le i \le n} \|\mathbf{l}_i^T\| \\ &\le n \cdot (\chi_A \|A\|)^{4a} \cdot \|Q^T B^{-T}\| \\ &= n \cdot (\chi_A \|A\|)^{4a} \cdot \|B^{-T}\| \\ &\le n \cdot \chi_A \cdot (\chi_A \|A\|)^{4a}, \end{aligned}$$

as claimed. Notice that the third line follows from the definition of C_1 and (4.8). The fourth line follows from the definition of L, and the last line uses (4.2). ∎

Next we show that the computed $\hat{R}$ in Step 1 is close to the true R using a forward error analysis.

THEOREM 4.6. *Let $\mathbf{r}_j^T$ and $\hat{\mathbf{r}}_j^T$ be the jth rows of R and $\hat{R}$, respectively. Then*

$$\frac{\|\mathbf{r}_j^T - \hat{\mathbf{r}}_j^T\|}{\|\mathbf{r}_j^T\|} \le c\epsilon \cdot \chi_A \|A\| \cdot \left\{ \sum_{i=j}^{m} \left[1 + (\chi_A \|A\|)^6 \right] \right\}^{1/2} + O(\epsilon^2), \tag{4.11}$$

where c is a small constant.

Remark. For this proof we assume that Householder transformations are used for the QR factorization in Step 1, although modified Gram–Schmidt or Givens rotations could also be used. In addition, we assume that the diagonal entries of R have the same signs as corresponding entries of $\hat{R}$. (Recall that QR factorization is uniquely determined only if an assumption is made about the signs of the diagonal entries of R.)

Proof. Let B be the $n \times n$ matrix of basis columns of A^T and consider first the factorization $B = QR_0$. Note that this Q is the same as Q in Step 1 and that R_0 denotes the first n columns of $RD^{1/2}$. Let $\hat{Q}$ and $\hat{R}_0$ denote the computed versions of these matrices. It follows from standard backward error analysis (see Theorem 18.4 of [15]) that there exists an exactly orthogonal matrix $\tilde{Q}$ such that $\|\tilde{Q} - \hat{Q}\| \leq c\epsilon$ and such that $\tilde{Q}\hat{R}_0 = QR_0 + E$, where $\|E\| \leq c\epsilon\|B\|$, where c is a small constant. We must change this to a forward error bound on the computed factors. If we multiply on the left by Q^T and the right by R_0^{-1}, we obtain $Q^T\tilde{Q}\hat{R}_0R_0^{-1} = I + Q^TER_0^{-1}$. Let $E' = Q^TER_0^{-1}$. By multiplying the preceding equation by its transpose, we obtain

$$(\hat{R}_0R_0^{-1})^T(\hat{R}_0R_0^{-1}) = (I + E')^T(I + E') = I + E' + (E')^T + (E')^TE'.$$

Notice that the left-hand side is a Cholesky factorization; i.e., $\hat{R}_0R_0^{-1}$ is an upper triangular matrix with positive diagonal entries. The Cholesky factorization is uniquely determined if the signs of diagonal entries are positive.

The right-hand side can be written as $I + E' + (E')^T + O(\epsilon^2)$. Write $E' + (E')^T = U^T + D + U$, where U is strictly upper triangular and D is diagonal. Then we observe that $(I + D/2 + U)^T(I + D/2 + U) = I + E' + (E')^T + O(\epsilon^2)$ and that $I + D/2 + U$ is upper triangular. Thus, by the uniqueness of the Cholesky factorization, we have $\hat{R}_0R_0^{-1} = I + D/2 + U + O(\epsilon^2)$, i.e.,

$$\hat{R}_0 = (I + D/2 + U)R_0 + O(\epsilon^2),$$

i.e.,

$$\hat{R}_0 - R_0 = (D/2 + U)R_0 + O(\epsilon^2).$$

Recall that $D/2 + U$ is part of $E' + (E')^T$. So

$$\|D/2 + U\| \leq 2\|E'\| \leq 2\|E\| \cdot \|R_0^{-1}\|.$$

Recall that $\|E\|$ is bounded by $c\epsilon\|B\|$ and $\|R_0^{-1}\| = \|B^{-1}\|$, $\|R_0\| = \|B\|$. Thus,

$$\|\hat{R}_0 - R_0\| \leq c\epsilon \cdot \|B\| \cdot \kappa(B) + O(\epsilon^2)$$

(with a different c). Recall from (4.2) that $\kappa(B) \leq \chi_A\|A\|$; hence

$$\|\hat{R}_0 - R_0\| \leq c\epsilon\|A\| \cdot (\chi_A\|A\|) + O(\epsilon^2).$$

A similar analysis, starting from the fact that $Q^T\tilde{Q} = \hat{R}_0R_0^{-1} + E' = I + D/2 + U + E' + O(\epsilon^2)$, gives a bound on $\|\hat{Q} - Q\|$ of $c\epsilon\chi_A\|A\|$. We will stop writing $O(\epsilon^2)$ for now.

Next we consider the actual QR factorization of $A^TD^{-1/2}$ in Step 1 of the COD algorithm. Since we have now proved that the computed Q is close to the exact Q, we conclude that the computed $\hat{R}$, which is $\hat{Q}A^TD^{-1/2}$, is close to the true R on a

column-by-column basis. In other words, let us define $\mathbf{v}_i$ to be the ith column of R and $\hat{\mathbf{v}}_i$ to be the corresponding column of $\hat{R}$; then

$$\|\mathbf{v}_i - \hat{\mathbf{v}}_i\| \leq c \cdot \epsilon \cdot \chi_A \|A\| \cdot \|\mathbf{v}_i\|, 1 \leq i \leq m.$$

This gives a bound on the elementwise error; namely,

$$|r_{ji} - \hat{r}_{ji}| \leq \|\mathbf{v}_i - \hat{\mathbf{v}}_i\| \leq c\epsilon\chi_A \|A\| \cdot \|\mathbf{v}_i\|, 1 \leq i \leq m, 1 \leq j \leq n.$$

Notice that if the ith row $\mathbf{a}_i^T$ of A is not linearly independent of the first $j-1$ basis rows, then $r_{ji} = 0$ if $i \geq j$. Because of the dependency test in the COD algorithm, $\hat{r}_{ji} = 0$ also. This point will be important later in the proof. We can now find a bound on the normwise error of the rows of R. Let $\mathbf{r}_j^T$ and $\hat{\mathbf{r}}_j^T$ be the jth rows of R and the computed matrix $\hat{R}$. Then

$$\frac{\|\mathbf{r}_j^T - \hat{\mathbf{r}}_j^T\|^2}{\|\mathbf{r}_j^T\|^2} \leq \frac{\sum_{i=j}^m (r_{ji} - \hat{r}_{ji})^2}{\sum_{i=j}^m r_{ji}^2} \leq \frac{c^2\epsilon^2(\chi_A\|A\|)^2 \cdot \sum_{i=j}^m \|\mathbf{v}_i\|^2}{r_{jj}^2}.$$

Based on the argument above, there will be no contribution (in the sum) from columns $\mathbf{v}_i$ if $\mathbf{a}_i^T$ is dependent only on the first $j-1$ basis rows. Now consider all other $\frac{\|\mathbf{v}_i\|^2}{r_{jj}^2}$. Suppose that $j-1, j \geq 1$ steps of the QR factorization have been completed. Let X denote $A^T D^{-1/2}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_m$ the columns of X. Then

$$Q_{j-1}^T \cdots Q_1^T A^T D^{-1/2} = \begin{bmatrix} r_{11} & \cdots & r_{1,j-1} & r_{1j} & \cdots & r_{1m} \\ 0 & & \vdots & \vdots & & \vdots \\ \vdots & & r_{j-1,j-1} & r_{j-1,j} & \cdots & r_{j-1,m} \\ \mathbf{0} & \cdots & \mathbf{0} & \bar{\mathbf{x}}_j & \cdots & \bar{\mathbf{x}}_m \end{bmatrix}.$$

Then $r_{jj}^2 = \|\bar{\mathbf{x}}_j\|^2$. Since no pivoting is necessary, the columns of A are ordered such that $\|\bar{\mathbf{x}}_i\|^2 \leq \|\bar{\mathbf{x}}_j\|^2$ for $i \geq j$. Recall that $\bar{R} = Q^T A^T = RD^{1/2}$. So for $i \geq j$,

$$\begin{aligned} \frac{\|\mathbf{v}_i\|^2}{r_{jj}^2} &= \frac{\|\bar{\mathbf{x}}_i\|^2 + r_{1i}^2 + \cdots + r_{j-1,i}^2}{r_{jj}^2} \\ &\leq 1 + \frac{r_{1i}^2 + \cdots + r_{j-1,i}^2}{r_{jj}^2} \\ &= 1 + \frac{d_i^{-1}\left(\bar{r}_{1i}^2 + \cdots + \bar{r}_{j-1,i}^2\right)}{d_j^{-1}\bar{r}_{jj}^2} \\ &\leq 1 + \frac{d_j\|\bar{\mathbf{r}}_i\|^2}{d_i\bar{r}_{jj}^2} \\ &\leq 1 + \chi_A^2 \cdot \|\mathbf{a}_i\|^2 \cdot (\chi_A\|A\|)^4 \\ &\leq 1 + (\chi_A\|A\|)^6, \end{aligned}$$

where $\bar{\mathbf{r}}_i$ in the fourth line is the ith column of $\bar{R}$. The fifth line follows from (4.4), (4.5), and the lower bound of (4.7) and holds only when $\mathbf{a}_i$ is linearly independent of $\mathbf{a}_1, \ldots, \mathbf{a}_{j-1}$. Recall that we need not be concerned with the other case since, as discussed above, there is no contribution from such terms. Thus,

$$\frac{\|\mathbf{r}_j^T - \hat{\mathbf{r}}_j^T\|^2}{\|\mathbf{r}_j^T\|^2} \leq c^2\epsilon^2(\chi_A\|A\|)^2 \cdot \left(\sum_{i=j}^m \left[1 + (\chi_A\|A\|)^6\right]\right)^2.$$

Taking the square root of both sides gives the required result. ∎

We have now established that the QR factorization in the first step of the algorithm gives an upper triangular matrix R that is well conditioned up to a scaling of the rows or the columns. It also yields a computed matrix $\hat{R}$ whose rows are close to those of R. With these results in hand, we move on to the analysis of the second step of the algorithm.

5. The second QR factorization. Recall that in this step we use the "skinny" QR factorization

$$R^T = Z_1 U_1,$$

where Z_1 is an $m \times n$ matrix with orthonormal columns and U_1 is an $n \times n$ upper triangular matrix. The results of section 4 imply that R^T is well conditioned up to a scaling of the columns. The QR factorization in this step, then, gives an upper triangular matrix $\hat{U}_1$ that is close to the exact upper triangular matrix U_1. In addition, the results (concerning the condition number of R) of section 4 can be used to prove similar results about the condition number of U_1. Again, let $D = \mathrm{diag}(d_1, \ldots, d_m)$ and $\bar{D} = D(1:n, 1:n)$. Since U_1 is the coefficient matrix of the system of equations in the backsubstitution step, the following theorem concerning the condition of U_1 will be quite useful.

THEOREM 5.1. *Let U_1 and $\bar{D}$ be defined as above. Then*

$$\kappa\left(\bar{D}^a U_1 \bar{D}^{1/2-a}\right) \leq n^{20} \cdot (\chi_A \|A\|)^{52} \tag{5.1}$$

for any $-\frac{1}{2} \leq a \leq \frac{1}{2}$.

Proof. We separately bound $\|\bar{D}^a U_1 \bar{D}^{1/2-a}\|$ and $\|\bar{D}^{a-1/2} U_1^{-1} \bar{D}^{-a}\|$, starting with the second norm. Notice that

$$R^T U_1^{-1} = Z_1.$$

Let $\mathbf{v}_k = U_1^{-1}(1:k, k)$, $\tilde{R} = R(1:k, :)$, and $\tilde{D} = D(1:k, 1:k)$. It follows from the fact that $U_1^{-T} R R^T U_1^{-1} = I$ that

$$\frac{1}{u_{kk}} \tilde{R}\tilde{R}^T \mathbf{v}_k = \mathbf{e}_k,$$

where $\mathbf{e}_k$ is the kth column of the $k \times k$ identity matrix. So

$$\begin{aligned}
\mathbf{v}_k &= u_{kk} \left(\tilde{R}\tilde{R}^T\right)^{-1} \mathbf{e}_k \\
&= \left(\mathbf{z}_k^T \mathbf{r}_k\right) \tilde{D}^{1/2} \left(\tilde{D}^{1/2} \tilde{R}\tilde{R}^T \tilde{D}^{1/2}\right)^{-1} \tilde{D}^{1/2} \mathbf{e}_k \\
&= \left(\mathbf{z}_k^T \mathbf{r}_k\right) d_k^{1/2} \tilde{D}^{1/2} \mathbf{x},
\end{aligned}$$

where $\mathbf{z}_k$ is the kth column of Z_1, $\mathbf{r}_k$ is the kth column of R^T, and $\mathbf{x}$ is the last column of $(\tilde{D}^{1/2}\tilde{R}\tilde{R}^T\tilde{D}^{1/2})^{-1}$. Multiplying both sides by $d_k^{-a}\tilde{D}^{a-1/2}$ yields

$$d_k^{-a} \tilde{D}^{a-1/2} \mathbf{v}_k = \left(\mathbf{z}_k^T \mathbf{r}_k\right) d_k^{1/2-a} \tilde{D}^a \mathbf{x}.$$

We show that there is an upper bound on the right-hand side as follows:

$$\| \left(\mathbf{z}_k^T \mathbf{r}_k\right) d_k^{1/2-a} \tilde{D}^a \mathbf{x}\| = d_k^{1/2} \cdot \left|\mathbf{z}_k^T \mathbf{r}_k\right| \cdot \|d_k^{-a} \tilde{D}^a \mathbf{x}\|$$

$$
\begin{aligned}
&\le d_k^{1/2} \cdot \|\mathbf{r}_k\| \cdot \|\tilde{D}^a (\tilde{D}^{1/2} \tilde{R} \tilde{R}^T \tilde{D}^{1/2})^{-1} \tilde{D}^{-a}\| \\
&\le \|A\| \cdot (\chi_A \|A\|)^2 \cdot \|(\tilde{D}^{1/2+a} \tilde{R} \tilde{R}^T \tilde{D}^{1/2-a})^{-1}\| \\
&= \|A\| \cdot (\chi_A \|A\|)^2 \cdot \frac{\kappa\left(\tilde{D}^{1/2+a} \tilde{R} \tilde{R}^T \tilde{D}^{1/2-a}\right)}{\|\tilde{D}^{1/2+a} \tilde{R} \tilde{R}^T \tilde{D}^{1/2-a}\|} \\
&\le \|A\| \cdot (\chi_A \|A\|)^2 \cdot \frac{\kappa\left(\tilde{D}^{1/2+a} \tilde{R} D^{-a} D^a \tilde{R}^T \tilde{D}^{1/2-a}\right)}{\bar{r}_{11}^2} \\
&\le \chi_A \cdot (\chi_A \|A\|)^3 \cdot \kappa\left(\tilde{D}^{1/2+a} \tilde{R} D^{-a}\right) \cdot \kappa\left(D^a \tilde{R}^T \tilde{D}^{1/2-a}\right).
\end{aligned}
$$

The third line is an application of (4.8) with $a = \frac{1}{2}$. In the fifth line of the above inequality, $\bar{r}_{11}$ is the (1,1) entry of $\bar{R} = Q^T A^T$. Notice that if $-\frac{1}{2} \le a \le \frac{1}{2}$, then Theorem 4.4 applies. So

$$
\begin{aligned}
\|d_k^{-a} \tilde{D}^{1/2-a} \mathbf{v}_k\| &\le \chi_A \cdot (\chi_A \|A\|)^3 \cdot \kappa\left(\tilde{D}^{1/2+a} \tilde{R} D^{-a}\right) \cdot \kappa\left(D^a \tilde{R}^T \tilde{D}^{1/2-a}\right) \\
&\le n^8 \cdot \chi_A \cdot (\chi_A \|A\|)^{23}.
\end{aligned}
$$

Now

$$
\begin{aligned}
\|\bar{D}^{a-1/2} U_1^{-1} \bar{D}^{-a}\| &\le \sum_{i=1}^{n} \|d_i^{-a} \bar{D}^{a-1/2} \mathbf{v}_i\| \\
&\le n \cdot \max_{1 \le i \le n} \|d_i^{-a} \tilde{D}^{a-1/2} \mathbf{v}_i\| \\
&\le n^9 \cdot \chi_A \cdot (\chi_A \|A\|)^{23} \qquad (5.2)
\end{aligned}
$$

for $-\frac{1}{2} \le a \le \frac{1}{2}$.

In order to find an upper bound on $\|\bar{D}^a U_1 \bar{D}^{1/2-a}\|$, we proceed as follows. First, recall that $Z_1^T Z_1 = I$ and that $Z_1 = R^T U_1^{-1}$. Combining these yields the identity $U_1^{-T} R R^T = U_1$. Thus,

$$
\begin{aligned}
\bar{D}^a U_1 \bar{D}^{1/2-a} &= \bar{D}^a U_1^{-T} R R^T \bar{D}^{1/2-a} \\
&= \bar{D}^a U_1^{-T} \bar{D}^{-a-1/2} \bar{D}^{a+1/2} R D^{-a} D^a R^T \bar{D}^{1/2-a},
\end{aligned}
$$

so

$$
\begin{aligned}
\|\bar{D}^a U_1 \bar{D}^{1/2-a}\| &\le \|\bar{D}^a U_1^{-T} \bar{D}^{-a-1/2}\| \cdot \|\bar{D}^{a+1/2} R D^a\| \cdot \|D^a R^T \bar{D}^a R^T \bar{D}^{1/2-a}\| \\
&\le n^9 \chi_A \, (\chi_A \|A\|)^{23} \cdot n \cdot \|A\| \cdot (\chi_A \|A\|)^{4(a+1/2)} \cdot n \cdot \|A\| \cdot (\chi_A \|A\|)^{4(1/2-a)}.
\end{aligned}
$$

The second line was obtained from the first by using (5.2) to bound the first factor on the right-hand side and (4.9) (with shifted values of "a") for the other two factors. The last line is written more simply as $n^{11} \cdot \|A\| \cdot (\chi_A \|A\|)^{28}$. Thus for $-\frac{1}{2} \le a \le \frac{1}{2}$,

$$
\begin{aligned}
\kappa\left(\bar{D}^a U_1 \bar{D}^{1/2-a}\right) &= \|\bar{D}^a U_1 \bar{D}^{1/2-a}\| \cdot \|\bar{D}^{a-1/2} U_1^{-1} \bar{D}^{-a}\| \\
&\le n^{20} \cdot (\chi_A \|A\|)^{52}. \qquad \square
\end{aligned}
$$

Now that we know that $\bar{D}^a U_1 \bar{D}^{1/2-a}$ is well conditioned for $-\frac{1}{2} \le a \le \frac{1}{2}$, we move on to the analysis of the remainder of the algorithm.

6. Finding the solution y. In analyzing the remainder of the algorithm, we first show that the error introduced in the backsubstitution step is small. In Step 3 of the algorithm, the upper triangular system

$$U_1\bar{\mathbf{y}} = Z_1^T D^{-1/2}\mathbf{b}$$

is solved for $\bar{\mathbf{y}}$. (Note that this is slightly different from the system given in Step 3 of the algorithm as presented in section 1 since the columns of $A^T D^{-1/2}$ have been "prepivoted.") Instead of working with the system given above, consider the following system:

$$\bar{D}^{1/2}U_1\bar{\mathbf{y}} = \bar{D}^{1/2}Z_1^T D^{-1/2}\mathbf{b},$$

where $D = \mathrm{diag}(d_1, d_2, \ldots, d_m)$ and $\bar{D} = D(1:n, 1:n)$ as before. In working through the steps of backsubstitution, one can see that solving this system is equivalent to solving the original one, even in floating-point arithmetic. (In other words, a rescaling of the rows does not change the numerical bounds.) Recall from the last section that $\bar{D}^a U_1 \bar{D}^{1/2-a}$ is well conditioned for $-\frac{1}{2} \le a \le \frac{1}{2}$. Therefore, standard techniques for analyzing backsubstitution can be used to show that the error at this step is small. The following theorem states that error bound.

THEOREM 6.1. *Let $\bar{\mathbf{y}}$ be the exact solution to $\bar{D}^{1/2}U_1\bar{\mathbf{y}} = \bar{D}^{1/2}Z_1^T D^{-1/2}\mathbf{b}$, and let $\check{\mathbf{y}}$ be the computed solution. Then*

$$\|\bar{\mathbf{y}} - \check{\mathbf{y}}\| \le \epsilon \cdot n^{39} \cdot \chi_A \cdot (\chi_A \|A\|)^{103} \cdot \|\mathbf{b}\| + \mathrm{O}(\epsilon^2). \tag{6.1}$$

Proof. Let $\check{\mathbf{y}}$ be the computed solution to the above system. Then $\check{\mathbf{y}}$ is the exact solution to the nearby system of equations

$$(\bar{D}^{1/2}U_1 + E)\check{\mathbf{y}} = \bar{D}^{1/2}Z_1^T D^{-1/2}\mathbf{b}.$$

The matrix E accounts for errors during the backsubstitution and $|E| \le \epsilon \cdot \left|\bar{D}^{1/2}U_1\right|$, where ϵ is machine roundoff [10]. So

$$\bar{D}^{1/2}U_1\bar{\mathbf{y}} - (\bar{D}^{1/2}U_1 + E)\check{\mathbf{y}} = \mathbf{0}$$

or

$$\bar{\mathbf{y}} - \check{\mathbf{y}} = (\bar{D}^{1/2}U_1)^{-1}E\check{\mathbf{y}}.$$

Substituting for $\check{\mathbf{y}}$ on the right-hand side yields

$$\bar{\mathbf{y}} - \check{\mathbf{y}} = (D^{1/2}U)^{-1}E(D^{1/2}U + E)^{-1}\bar{D}^{1/2}Z_1^T D^{-1/2}\mathbf{b}.$$

Thus,

$$\begin{aligned}
\|\bar{\mathbf{y}} - \check{\mathbf{y}}\| &\le \|(\bar{D}^{1/2}U_1)^{-1}\| \cdot \|E\| \cdot \|(\bar{D}^{1/2}U_1 + E)^{-1}\| \cdot \|\bar{D}^{1/2}Z_1^T D^{-1/2}\| \cdot \|\mathbf{b}\| \\
&\le \epsilon \cdot \|(\bar{D}^{1/2}U_1)^{-1}\| \cdot \|\bar{D}^{1/2}U_1\| \cdot \|(\bar{D}^{1/2}U_1 + E)^{-1}\| \cdot \|\bar{D}^{1/2}U_1^{-T}RD^{-1/2}\| \cdot \|\mathbf{b}\| \\
&\le \epsilon \cdot \|(\bar{D}^{1/2}U_1)^{-1}\| \cdot \|\bar{D}^{1/2}U_1\| \cdot (\|I\| + \|(\bar{D}^{1/2}U_1)^{-1}E\| + \|(\bar{D}^{1/2}U_1)^{-1}E\|^2 \\
&\quad + \|(\bar{D}^{1/2}U_1)^{-1}E\|^3 + \cdots) \cdot \|(\bar{D}^{1/2}U_1)^{-1}\| \cdot \|\bar{D}^{1/2}U_1^{-T}\bar{D}^{-1}\bar{D}RD^{-1/2}\| \cdot \|\mathbf{b}\| \\
&\le \epsilon \cdot \kappa(\bar{D}^{1/2}U_1) \cdot \|(\bar{D}^{1/2}U_1)^{-1}\| \cdot \|\bar{D}^{1/2}U_1^{-T}\bar{D}^{-1}\| \cdot \|\bar{D}RD^{-1/2}\| \cdot \|\mathbf{b}\| + \mathrm{O}(\epsilon^2) \\
&\le \epsilon \cdot n^{39} \cdot \chi_A \cdot (\chi_A\|A\|)^{103} \cdot \|\mathbf{b}\| + \mathrm{O}(\epsilon^2),
\end{aligned}$$

as claimed. The last line is obtained by applying (5.1), (5.2), and (4.9) with the appropriate values of a. □

In the theorem above, the errors in the computation of U_1 itself (which also contribute to the error in $\bar{\mathbf{y}}$) are not included, but could be accounted for as a somewhat larger perturbation matrix E. As we have already argued in the proof of Theorem 4.6, the errors in computing the factors are small. A similar analysis could be applied to the second factorization, showing that errors made in forming each row of U_1 are small with respect to the norm of that row. Therefore, the perturbation matrix E is small with respect to $\bar{D}^{1/2}U_1$. Explicitly including this analysis in the previous theorem would make the proof more complicated, but the bound would be qualitatively the same.

The final step is to obtain $\mathbf{y}$ by multiplying $\bar{\mathbf{y}}$ by Q. Let $\hat{\mathbf{y}}$ be the computed result. Assume that $\hat{\mathbf{y}}$ accounts for the errors during both this step and the previous step. Equations (5.2), (4.9), and (6.1) are used to obtain the following error bound:

$$\begin{aligned}\|\mathbf{y}-\hat{\mathbf{y}}\| &\le \epsilon\cdot n\cdot\|\mathbf{y}\| + \|\bar{\mathbf{y}}-\check{\mathbf{y}}\| \\ &\le \epsilon\cdot n\cdot\|(\bar{D}^{1/2}U_1)^{-1}\|\cdot\|\bar{D}^{-1}U_1^{-1}\bar{D}^{1/2}\|\cdot\|\bar{D}RD^{-1/2}\|\cdot\|\mathbf{b}\| + \|\bar{\mathbf{y}}-\check{\mathbf{y}}\| \\ &\le \epsilon\cdot\left[n^{19}\cdot\chi_A\cdot(\chi_A\|A\|)^{51} + n^{39}\cdot\chi_A\cdot(\chi_A\|A\|)^{103}\right]\cdot\|\mathbf{b}\| + \mathrm{O}(\epsilon^2). \qquad (6.2)\end{aligned}$$

Notice that the error bound is of the form

$$\|\mathbf{y}-\hat{\mathbf{y}}\| \le \epsilon\cdot f(A)\cdot\|\mathbf{b}\|.$$

Thus, the COD algorithm satisfies the definition of stability.

7. Related work. As discussed in section 2, weighted least-squares problems arise in a number of applications. Consequently, there are numerous algorithms in the literature that are specialized for weighted least squares. In this section, we give a brief overview of backward error analysis, which is used in the stability analyses of these algorithms, and describe the relationship to the forward error analysis of this paper. In doing so, we show that there are difficulties in trying to obtain stability bound (1.2) in such a setting, and we explain how the COD algorithm and its analysis avoid these problems. We also take a closer look at several algorithms for which a forward analysis has been done or for which proving a forward bound appears possible.

As just mentioned, there are many algorithms that appear in the literature. Such algorithms are presented in Barlow [1], Björck and Duff [2], Golub [9], Gulliksson [13], Gulliksson and Wedin [14], Paige [20], Peters and Wilkinson [21], Powell and Reid [22], Van Loan [26], and Vavasis [28]. These algorithms are based on standard algorithms for solving unweighted least-squares problems, and special techniques are employed to exploit structure and to deal with widely varying weights. None of these works, except [28], proves a forward stability bound for their algorithms. Many of these papers were published before Theorem 1.1 appeared, so the absence of a forward error bound is not surprising. Recall that a forward error bound has relevance for the applications described in section 2.

Since Vavasis's NSH method [28] is the only other algorithm that proves a forward stability bound, we discuss it first. The NSH algorithm employs nonstandard techniques, particularly when choosing the null space basis for A^TD^{-1}. In contrast, the COD algorithm of this paper uses standard techniques that are well understood, namely, QR decomposition and backsubstitution. Also, our algorithm is more efficient than the NSH algorithm. The NSH method solves an $m\times m$ system of equations and

thus requires $O(m^3)$ flops. The work for the QR factorizations dominates the work required for the complete orthogonal algorithm, so this algorithm requires $O(mn^2)$ flops. Since $n < m$ (and n could be much smaller than m), the COD algorithm requires less work.

In considering the other algorithms, the question arises whether a forward error bound can be derived from the error analyses given by some of the authors mentioned above. Several (e.g., [1], [2], [13], [20], [22]) prove a backward stability bound; i.e., they prove that their algorithms compute a solution $\hat{\mathbf{y}}$ that solves a nearby problem given by

$$\min_{\hat{\mathbf{y}} \in \mathbb{R}^n} \| (D + \Delta D)^{-1/2} [(A + \Delta A)\, \hat{\mathbf{y}} - (\mathbf{b} + \Delta \mathbf{b})] \|$$

(or something similar), where $\|\Delta D\| \leq \epsilon \cdot c_1 \cdot \|D\|$, $\|\Delta A\| \leq \epsilon \cdot c_2 \cdot \|A\|$, and $\|\Delta \mathbf{b}\| \leq \epsilon \cdot c_3 \cdot \|\mathbf{b}\|$. Here c_1, c_2, and c_3 are small constants and ϵ is the machine precision. It may seem at first that a forward error bound can be obtained by applying Theorem 1.1 to the backward error bounds. Since all of the backward error analyses of previous authors introduce a perturbation ΔA into the coefficient matrix A, the forward error bound will involve $\chi_{A+\Delta A}$. The result is a forward bound of the form

$$\|\mathbf{y} - \hat{\mathbf{y}}\| \leq \epsilon \cdot c \cdot \chi_{A+\Delta A} \cdot \|\mathbf{b}\|, \tag{7.1}$$

where c is a constant for these other algorithms. Here a difficulty arises: χ_A is not continuous with respect to perturbations of A as observed by [23]. In fact, there exists a matrix A such that $\chi_A < 3$ but for any $\epsilon > 0$

$$\sup\{\chi_{A+\Delta A} : \|\Delta A\| \leq \epsilon\} = \infty.$$

In particular, the example is

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

Note that $\lim_{n\to\infty} \chi_{A_n} = \infty$ if we define

$$A_n = \begin{bmatrix} 1 + 1/n & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

Thus, the supremum of the right-hand side of (7.1) over all arbitrarily small perturbations of A is infinity for this particular A. Notice that the matrix A has the special property that its first two rows are parallel. Indeed, it is a consequence of [23] and [19] that χ_A is discontinuous at A whenever A has an $n \times n$ singular submatrix. (This fact also follows from (4.2).) A "randomly chosen" matrix would never have an $n \times n$ singular submatrix, but singular submatrices occur often in practice (e.g., consider a node-arc adjacency matrix or a linear programming problem that includes two simple bound constraints of the form $y_i \geq c_1$ and $y_i \leq c_2$). Thus, (1.2) is not established for such an A; i.e., it cannot be established in general using the kind of backward error analysis in the literature mentioned.

The most difficult question with respect to previous algorithms is the following: Can a forward error bound be derived for such an algorithm by using a completely new

analysis? We claim that any algorithm purporting to satisfy a forward error bound like (1.2) must have an explicit test for singularity. To illustrate this, we consider the impact of a singular submatrix upon algorithms for weighted least squares. Consider the following two problems:

$$\min \left\| \operatorname{diag}(10^{30}, 10^{30}, 1) \left(\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \mathbf{y} - \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \right) \right\| \tag{7.2}$$

and

$$\min \left\| \operatorname{diag}(10^{30}, 10^{30}, 1) \left(\begin{bmatrix} 1 + 10^{-15} & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \mathbf{y} - \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \right) \right\|. \tag{7.3}$$

Observe that although the numerical data in (7.2) is very close to (7.3), the two problems have sharply different solutions. In (7.2), the first two rows are parallel and are minimized (in the absence of the third row) by any vector on the line $y_1 + y_2 = 1.5$. The third row is minimized by any vector with $y_2 = 3$. Thus, the solution to the problem is $(-1.5, 3)$.

In contrast, in (7.3) the first two rows are not parallel. Because of their huge weights, both must be satisfied nearly as equations at the weighted least-squares solution; i.e., the solution to the second problem will be very close to

$$\begin{bmatrix} 1 + 10^{-15} & 1 \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 2 \end{bmatrix},$$

which will obviously be very large since the matrix being inverted is nearly singular.

Thus, any algorithm purporting to satisfy a forward error bound like (1.2) for the example problem (7.2) must preserve the dependence between the first two rows of A even if there is roundoff error. In fact, this is exactly how COD operates. The COD algorithm, when it encounters rows of A that are nearly dependent, will perturb them so that they become exactly dependent. This perturbation is the topic of section 3. Ordinarily, it is considered undesirable for a numerical algorithm to make a yes/no decision about linear dependence. It is apparent from the contrast between (7.2) and (7.3), however, that any algorithm that is supposed to satisfy stability bound (1.2) must "know" that the first two rows of A in (7.2) are parallel.

Thus, we conclude that forward-stable algorithms, that is, algorithms satisfying (1.2), must apparently include a test for linear dependence among the rows with heaviest weights. Most algorithms in the literature do not. One exception is Björck and Duff's modification of Peters and Wilkinson's algorithm. This method is an elimination-based decomposition of the normal equations. A special type of pivoting is used to preserve sparsity. We suspect that a modification of this pivoting and an appropriate test for linear dependence among the rows of A will yield an algorithm for which a forward stability bound can be established. The only approach we know of to establish this forward error bound would be an analysis similar to the preceding analysis of COD.

Additionally, many interior-point implementations of Cholesky factorization effectively perform a test for linear dependence, e.g., [32]. If a Cholesky pivot is very small, then the algorithm treats that column differently. Although a singularity test is present, we do not expect that a method based on normal equations could be stable

for weighted least squares. Too much information about rows of A corresponding to small entries of D^{-1} is lost when forming $A^T D^{-1} A$ (e.g., consider what happens to the last row of A in the normal equations of (7.2)). Nevertheless, interior-point practitioners report success with this approach.

8. Summary and open questions. The weighted least-squares problem

$$\min_{\mathbf{y} \in \mathbb{R}^n} \| D^{-1/2} (A\mathbf{y} - \mathbf{b}) \|,$$

where $D \in \mathbb{R}^{m \times m}$ is a diagonal, positive-definite, ill-conditioned matrix; $A \in \mathbb{R}^{m \times n}$ is a full-rank matrix; $\mathbf{y} \in \mathbb{R}^n$; and $\mathbf{b} \in \mathbb{R}^m$, has a unique solution. Because of the ill conditioning of D, the standard methods for solving least-squares problems do not find an accurate solution. We have employed a version of COD for this problem. The COD algorithm involves four steps, given in section 1. We then proceeded to show that this algorithm is stable, as defined in section 1. The only previously known stable algorithm is from [28], but it is much more complicated and also requires more flops than the COD method. Other previous algorithms do not satisfy a forward error bound of the form we seek.

Now that we know that the algorithm is stable, there are several open questions.

1. This paper contains a forward error analysis of the COD algorithm. The final bound, which is (6.2), involves some very large factors. These factors appear to be an artifact of our forward analysis rather than a feature of the algorithm. The alternative to forward error analysis is backward error analysis. For many other problems in numerical linear algebra, backward error analysis is successful in ultimately producing the best known forward error bounds [10].

As pointed out in section 7, the straightforward approach to backward error analysis for weighted least squares could not yield the kind of forward error bound we seek. Is it possible to do a specialized backward error analysis of this algorithm, and will such an analysis yield better forward bounds? It appears that such a backward error analysis must be restricted to some special class of structured perturbations to A.

2. This algorithm has been implemented using dense methods. In many applications, the matrix A is sparse. Can we implement this algorithm in such a way that it takes advantage of that sparsity?

3. The results thus far are theoretical. This algorithm has not yet been tested in larger applications. The question, then, is whether or not this algorithm is effective in applications. We are currently beginning tests of our algorithm in interior-point methods [16].

The problem of stably solving the ill-conditioned equilibrium system in barrier methods for optimization has received a fair amount of attention [7]. In the case of barrier methods for linear programming (that is, interior-point methods), the equilibrium system reduces to weighted least squares, which is the problem addressed by this paper. Other authors have recently looked at ill conditioning in barrier methods, including Coleman and Liu [4], Forsgren, Gill, and Shinnerl [6], Gill, Saunders, and Shinnerl [8], Gould [12], Murray [17], Nash and Sofer [18], M. H. Wright [29], and S. J. Wright [30].

One difference between these other works and ours may be summarized as follows. These other works typically look at the more general problem $\min \|H^{-1/2}(A\mathbf{y} - \mathbf{b})\|$, where H is symmetric and positive definite, but not necessarily diagonal. This is a problem that we currently cannot address with our techniques. In some recent work,

Forsgren [5] derived a result similar to Theorem 1.1 for such matrices H that are also diagonally dominant. The applicability of the COD algorithm to this case has not yet been determined.

On the other hand, when specialized to diagonal weight matrices D, these authors consider a more restricted problem in that they all make an assumption that the large and small entries on the diagonal of D have some correlation with the columns of A^T. This corresponds to a nondegeneracy assumption about the underlying optimization problem. In contrast, our method does not involve any restrictions about where "large" versus "small" entries of D can appear, and thus it is hoped that the COD method has less difficulty when there is degeneracy or near-degeneracy in the underlying optimization problem.

The final bounds proved by some of these authors—[29] and [30] in particular, but also [6], [12], and [18]—specifically address the following issue for interior-point methods for the nondegenerate case. For an interior-point method, the steps $\Delta\mathbf{x}, \Delta\mathbf{s}$ are small in some components as convergence is approached because of a special correlation between the weight matrix and the right-hand side of (2.1), as mentioned in section 2. This means that a better bound than (1.2) is needed for interior-point methods. Our present analysis is not sufficiently strong to address this issue. This will be considered in a forthcoming paper [16], and computational tests will be presented.

4. The COD algorithm is a direct method for solving the weighted least-squares problem. Another approach to solving this problem is to solve the normal equations using iterative methods. This approach is currently being investigated by Bobrovnikova and Vavasis.

Acknowledgments. We revised this paper taking into account the helpful comments of two anonymous referees and the journal editor.

REFERENCES

[1] J. L. Barlow, *Stability analysis of the G-algorithm and a note on its application to sparse least squares problems*, BIT, 25 (1985), pp. 507–520.

[2] A. Björck and I. S. Duff, *A direct method for the solution of sparse linear least squares problems*, Linear Algebra Appl., 34 (1980), pp. 43–67.

[3] L. O. Chua, C. A. Desoer, and E. S. Kuh, *Linear and Nonlinear Circuits*, McGraw–Hill, New York, 1987.

[4] T. F. Coleman and J. Liu, *An Interior Newton Method for Quadratic Programming*, Tech. report CTC93TR153, Advanced Computing Research Institute, Cornell University, Ithaca, NY, 1993.

[5] A. L. Forsgren, *On Linear Least-Squares Problems with Diagonally Dominant Weight Matrices*, Report TRITA-MAT-1995-OS2, Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden, 1995.

[6] A. Forsgren, P. E. Gill, and J. R. Shinnerl, *Stability of symmetric ill-conditioned systems arising in interior methods for constrained optimization*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 187–211.

[7] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, London, 1981.

[8] P. E. Gill, M. A. Saunders, and J. R. Shinnerl, *On the stability of Cholesky factorization for symmetric quasidefinite systems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 35–46.

[9] G. Golub, *Numerical methods for solving linear least squares problems*, Numer. Math., 7 (1965), pp. 206–216.

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

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

[12] N. Gould, *On the accurate determination of search directions for simple differentiable penalty functions*, IMA J. Numer. Anal., 6 (1986), pp. 357–372.
[13] M. Gulliksson, *Backward error analysis for the constrained and weighted linear least squares problem when using the weighted QR factorization*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 675–687.
[14] M. Gulliksson and P. A. Wedin, *Modifying the QR decomposition to constrained and weighted linear least squares*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1298–1313.
[15] N. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996.
[16] P. D. Hough, *Stable and Efficient Solution of Weighted Least-Squares Problems with Applications in Interior Point Methods*, Ph.D. thesis, Cornell University, Ithaca, NY, 1996.
[17] W. Murray, *Analytical expressions for the eigenvalues and eigenvectors of the Hessian matrices of barrier and penalty functions*, J. Optim. Theory Appl., 7 (1971), pp. 189–196.
[18] S. G. Nash and A. Sofer, *A barrier method for large-scale constrained optimization*, ORSA J. Comput., 5 (1993), pp. 40–53.
[19] D. P. O'Leary, *On bounds for scaled projections and pseudoinverses*, Linear Algebra Appl., 132 (1990), pp. 115–117.
[20] C. C. Paige, *Fast numerically stable computations for generalized least squares problems*, SIAM J. Numer. Anal., 16 (1979), pp. 165–171.
[21] G. Peters and J. H. Wilkinson, *The least squares problem and pseudo-inverses*, Comput. J., 13 (1970), pp. 309–316.
[22] M. J. D. Powell and J. K. Reid, *On applying Householder transformations to linear least squares problems*, in Proc. IFIP Congress, 1968, North–Holland, Amsterdam, 1969, pp. 122–126.
[23] G. W. Stewart, *On scaled projections and pseudoinverses*, Linear Algebra Appl., 112 (1989), pp. 189–193.
[24] G. Strang, *A framework for equilibrium equations*, SIAM Rev., 30 (1988), pp. 283–297.
[25] M. J. Todd, *A Dantzig–Wolfe-like variant of Karmarkar's interior-point linear programming algorithm*, Oper. Res., 38 (1990), pp. 1006–1018.
[26] C. F. Van Loan, *On the method of weighting for equality-constrained least-squares problems*, SIAM J. Numer. Anal., 22 (1985), pp. 851–864.
[27] S. A. Vavasis, *Stable finite elements for problems with wild coefficients*, SIAM J. Numer. Anal., 33 (1996), pp. 890–916.
[28] S. A. Vavasis, *Stable numerical algorithms for equilibrium systems*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1108–1131.
[29] M. H. Wright, *Determining Subspace Information from the Hessian of a Barrier Function*, Tech. report 92-02, AT&T Bell Laboratories Numerical Analysis manuscript, 1992.
[30] S. J. Wright, *Stability of Linear Algebra Computations in Interior-Point Methods for Linear Programming*, Tech. report MCS-P446-0694, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 1994.
[31] S. J. Wright, *Primal-dual Interior-point Methods*, SIAM, 1997, to appear.
[32] Y. Zhang, *Solving Large-scale Linear Programs by Interior-Point Methods under the MATLAB Environment*, Tech. report TR96-01, Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD, 1996.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, pp. 393–402, April 1997

010

OPTIMAL BACKWARD PERTURBATION BOUNDS FOR UNDERDETERMINED SYSTEMS*

JI-GUANG SUN† AND ZHENG SUN‡

Abstract. Let A be an $n \times m\,(m < n)$ matrix, b be an m-dimensional vector, and y be an n-dimensional nonzero vector. In this paper we consider the following open problem: find an explicit expression of the optimal backward perturbation bound $\eta_\theta(y)$ defined by

$$\eta_\theta(y) = \inf\{\|(F^T, \theta g)\|_F : y \text{ is the minimum 2-norm solution to } (A+F)^T x = b + g\},$$

where θ is a positive number. This problem is solved.

Key words. underdetermined system, the minimum 2-norm solution, backward perturbation bound

AMS subject classifications. 15A06, 65F99

PII. S0895479896297434

1. Introduction.

Let $A^T x = b$ be an underdetermined system, where $A \in \mathcal{R}^{n\times m}$ with $m < n$. It is known [2] that the system either has no solution or has an infinity of solutions. Effective numerical methods and perturbation results can be found in the literature (see, e.g., [1], [2], [4], [7]).

Let $A \in \mathcal{R}^{n\times m}$ $(m < n)$, $b \in \mathcal{R}^m$, and $y \in \mathcal{R}^n$ be given. In this paper we consider the following open problem [3], [4, Problem 20.2]: find an explicit expression of the optimal backward perturbation bound $\eta_\theta(y)$ defined by

(1.1)
$$\eta_\theta(y) = \inf\left\{\|(F^T, \theta g)\|_F : y \text{ is the minimum 2-norm solution to } (A+F)^T x = b + g\right\},$$

where θ is a positive number. Note that the optimal backward perturbation bound $\eta_\theta(y)$ can also be called the normwise backward error [3], [4], [5].

For deriving an explicit expression of $\eta_\theta(y)$, we first consider a special case—only the coefficient matrix A is perturbed. Let $\mathcal{F}$ be the set defined by

$$\mathcal{F} = \{F \in \mathcal{R}^{n\times m} : y \text{ is the minimum 2-norm solution to } (A+F)^T x = b\}. \tag{1.2}$$

Obviously, $\mathcal{F}$ is the set of all backward perturbations F of A with respect to y and b. The optimal backward perturbation bound $\eta(y)$ is defined by

$$\eta(y) = \inf_{F\in\mathcal{F}} \|F\|_F. \tag{1.3}$$

Let $\mathcal{F}$ be the set defined by (1.2). Observe the following facts: (i) if $b = 0$ but $y \neq 0$, or if $y = 0$ but $b \neq 0$, then $\mathcal{F} = \emptyset$ (the empty set), (ii) if $b = 0$ and $y = 0$, then $\mathcal{F} = \mathcal{R}^{n\times m}$, and in such a case we have $\eta(y) = 0$. Hence, we shall assume $b \neq 0$ and $y \neq 0$ for the set $\mathcal{F}$.

* Received by the editors January 16, 1996; accepted for publication (in revised form) by N. J. Higham May 17, 1996.

http://www.siam.org/journals/simax/18-2/29743.html

† Department of Computing Science, Umeå University, S-901 87 Umeå, Sweden (jisun@cs.umu.se). The work of this author was supported by the Swedish Natural Science Research Council under contract M-AA/MA 06952-303 and the Department of Computing Science, Umeå University.

‡ Department of Electrical Engineering, Linköping University, S-581 83 Linköping, Sweden. This author is an undergraduate student. The greater part of this work was performed while this author was visiting Umeå during Christmas 1995 and New Year's Day 1996.

Some basic properties of the set $\mathcal{F}$ defined by (1.2) are revealed in section 2. An equivalent definition of $\mathcal{F}$ and a characterization of the closure of $\mathcal{F}$ are given. On the basis of the results of section 2, we derive an explicit expression of the optimal bound $\eta(y)$ in section 3, and the expression of $\eta(y)$ is then used to derive an explicit expression of the optimal bound $\eta_\theta(y)$ in section 4. Finally, in section 5 we give some remarks.

Throughout this paper we use the following notation. $A^\dagger$ denotes the Moore–Penrose inverse of a matrix A, $P_A = AA^\dagger$ denotes the orthogonal projection onto the column space $\mathcal{R}(A)$, and $P_A^\perp = I - P_A$. $\sigma_m(A)$ stands for the mth largest singular value of A and $\lambda_m(B)$ for the mth largest eigenvalue of a square matrix B having only real eigenvalues. $\|\cdot\|_2$ denotes the Euclidean vector norm and $\|\cdot\|_F$ the Frobenius norm. For a subset $\mathcal{S}$ of $\mathcal{R}^n$, the symbol $\overline{\mathcal{S}}$ denotes the closure of $\mathcal{S}$.

Before the main body of the paper, we now consider the simplest case: $m = 1$. Let $a \in \mathcal{R}^n$, nonzero $b \in \mathcal{R}$, and nonzero $y \in \mathcal{R}^n$ be given. By (1.2), the set $\mathcal{F}$ is defined by

$$\mathcal{F} = \{f \in \mathcal{R}^n : y \text{ is the minimum 2-norm solution to } (a+f)^T x = b\}.$$

It is easy to see that $\mathcal{F}$ contains a unique vector f, and the vector f satisfies

$$y = (a+f)^{T\dagger} b. \tag{1.4}$$

From (1.4) we get

$$f = \frac{by}{\|y\|_2^2} - a$$

and

$$\|f\|_2^2 = \frac{b^2 - 2ba^T y + \|a\|_2^2\|y\|_2^2}{\|y\|_2^2} = \frac{(b - a^T y)^2}{\|y\|_2^2} + \|(I - yy^\dagger)a\|_2^2.$$

Consequently,

$$\eta(y) = \sqrt{\frac{r^2}{\|y\|_2^2} + \sigma_1^2\left((I - yy^\dagger)a\right)}, \tag{1.5}$$

where $\sigma_1((I - yy^\dagger)a) = \|(I - yy^\dagger)a\|_2$ denotes the singular value of $(I - yy^\dagger)a$, and $r = b - a^T y$.

2. Lemmas. In this section we study basic properties of the set $\mathcal{F}$ defined by (1.2). The following result gives an equivalent definition of $\mathcal{F}$. A proof of the result can be found in [7, section 1].

LEMMA 2.1 [7], [1]. *Let $A \in \mathcal{R}^{n\times m}$, nonzero $b \in \mathcal{R}^m$, and nonzero $y \in \mathcal{R}^n$ be given. Let $\mathcal{F}$ be the set defined by* (1.2)*, and let $\mathcal{F}_1$ be the set defined by*

$$\mathcal{F}_1 = \{F \in \mathcal{R}^{n\times m} : (A+F)^T y = b \text{ and } y \in \mathcal{R}(A+F)\}. \tag{2.1}$$

Then $\mathcal{F}_1 = \mathcal{F}$.

It is worth pointing out that the set $\mathcal{F}\,(= \mathcal{F}_1)$ is not necessarily closed when $m > 1$. We now explain this fact by a simple example. Let

$$A = (a_1, a_2, a_3) = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & -1 & 0 \end{pmatrix}, \quad y = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad b = \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}.$$

Consider the matrices F_τ expressed by

$$F_\tau = (f_1^{(\tau)}, f_2^{(\tau)}, f_3^{(\tau)}) = \begin{pmatrix} 1 & -1 & 0 \\ 1 & -1+\tau & 0 \\ 0 & 2 & 0 \\ 0 & 2 & 1 \end{pmatrix} \quad \text{with} \quad \tau > 0.$$

From

$$A + F_\tau = \begin{pmatrix} 2 & 0 & 0 \\ 2 & \tau & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

we see that the matrices F_τ satisfy

$$(A + F_\tau)^T y = b$$

and

$$y = \frac{1}{2}(a_1 + f_1^{(\tau)}) - \frac{1}{\tau}(a_2 + f_2^{(\tau)}) + \frac{1}{\tau}(a_3 + f_3^{(\tau)}); \quad \text{i.e.,} \quad y \in \mathcal{R}(A + F_\tau).$$

Consequently, by (2.1) and Lemma 2.1, $F_\tau \in \mathcal{F}$ for any $\tau > 0$. Taking $\tau \to 0$, we get

$$\lim_{\tau \to 0} F_\tau = \begin{pmatrix} 1 & -1 & 0 \\ 1 & -1 & 0 \\ 0 & 2 & 0 \\ 0 & 2 & 1 \end{pmatrix} = F$$

and

$$A + F = \begin{pmatrix} 2 & 0 & 0 \\ 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

Obviously, the matrix F satisfies $(A+F)^T y = b$. However, $y \notin \mathcal{R}(A+F)$. Therefore, $F \notin \mathcal{F}$. This means that the limit point F of the matrices $F_\tau \in \mathcal{F}$ is not contained in $\mathcal{F}$. Hence, the set $\mathcal{F}$ associated with this example is not closed.

We now consider a special case:

$$y = (\eta_1, 0, \ldots, 0)^T \in \mathcal{R}^n, \quad b = (\beta_1, 0, \ldots, 0)^T \in \mathcal{R}^m, \quad \eta_1, \beta_1 > 0. \tag{2.2}$$

The following lemma presents a characterization of the closure $\overline{\mathcal{F}} (= \overline{\mathcal{F}}_1)$ of $\mathcal{F}$.

LEMMA 2.2. *Let $A, b, y, \mathcal{F}, \mathcal{F}_1$ be as in Lemma 2.1 with $m > 1$, and let b, y be in the forms of* (2.2). *Define the set $\mathcal{F}_2$ by*

$$\mathcal{F}_2 = \left\{ F = \begin{pmatrix} \beta_1/\eta_1 & 0 \\ z_1 & Z_2 \end{pmatrix} - A : \begin{array}{l} z_1 \in \mathcal{R}^{n-1}, Z_2 \in \mathcal{R}^{(n-1)\times(m-1)}, \\ \operatorname{rank}(z_1, Z_2) \le m-1 \end{array} \right\}. \tag{2.3}$$

Then $\mathcal{F}_2 = \overline{\mathcal{F}}_1$; i.e., $\mathcal{F}_2 = \overline{\mathcal{F}}$.

Proof. We only need to show the following three facts: (i) $\mathcal{F}_1 \subset \mathcal{F}_2$, (ii) $\mathcal{F}_2$ is a closed set, (iii) $\mathcal{F}_2 \subset \overline{\mathcal{F}}_1$.

Assume $F \in \mathcal{F}_1$. By (2.1) and (2.2), F satisfies

$$A + F = \begin{pmatrix} \beta_1/\eta_1 & 0 \\ z_1 & Z_2 \end{pmatrix} \quad \text{and} \quad z_1 \in \mathcal{R}(Z_2),$$

which implies $F \in \mathcal{F}_2$. Consequently, $\mathcal{F}_1 \subset \mathcal{F}_2$.

By (2.3), $\mathcal{F}_2$ is obviously a closed set.

Finally, we prove $\mathcal{F}_2 \subset \overline{\mathcal{F}}_1$. Assume $F \in \mathcal{F}_2$. By (2.3) we have $(A+F)^T y = b$. Let z_1, Z_2 be as in (2.3) and $\operatorname{rank}(z_1, Z_2) = j$. If $j = 0$, then from (2.3) it follows that $y \in \mathcal{R}(A+F)$. By (2.1) we have $F \in \mathcal{F}_1$. We now consider the case $1 \le j \le m-1$. By using the QR factorization or the column pivoted QR factorization [2, Chapters 5, 5.2, and 5.4.1], the matrix (z_1, Z_2) can be expressed by $(z_1, Z_2) = Q(c_1, C_2)$, where $Q \in \mathcal{R}^{(n-1)\times j}$, the columns of Q form an orthonormal basis of $\mathcal{R}(z_1, Z_2)$, and $c_1 \in \mathcal{R}^j, C_2 \in \mathcal{R}^{j\times(m-1)}, \operatorname{rank}(c_1, C_2) = j$. Consequently, in this case $A + F$ can be expressed by

$$A + F = \begin{pmatrix} \eta_1 & 0 \\ 0 & Q \end{pmatrix} \begin{pmatrix} \beta_1/\eta_1^2 & 0 \\ c_1 & C_2 \end{pmatrix}. \tag{2.4}$$

If $\operatorname{rank}(C_2) = j$, then from (2.4)

$$\begin{pmatrix} \eta_1 & 0 \\ 0 & Q \end{pmatrix} = (A + F) \begin{pmatrix} \beta_1/\eta_1^2 & 0 \\ c_1 & C_2 \end{pmatrix}^{\dagger},$$

which shows that $y = \binom{\eta_1}{0} \in \mathcal{R}(A+F)$. By (2.1) we have $F \in \mathcal{F}_1$. If $\operatorname{rank}(C_2) < j$, we can take a sequence $E_2^{(l)} \in \mathcal{R}^{j\times(m-1)}$ $(l = 1, 2, \ldots)$ that

$$\operatorname{rank}(C_2 + E_2^{(l)}) = j \quad \text{and} \quad \lim_{l\to\infty} E_2^{(l)} = 0. \tag{2.5}$$

Let

$$F^{(l)} = \begin{pmatrix} \eta_1 & 0 \\ 0 & Q \end{pmatrix} \begin{pmatrix} \beta_1/\eta_1^2 & 0 \\ c_1 & C_2 + E_2^{(l)} \end{pmatrix} - A, \quad l = 1, 2, \ldots. \tag{2.6}$$

Then $(A + F^{(l)})^T y = b$, and from (2.6)

$$\begin{pmatrix} \eta_1 & 0 \\ 0 & Q \end{pmatrix} = (A + F^{(l)}) \begin{pmatrix} \beta_1/\eta_1^2 & 0 \\ c_1 & C_2 + E_2^{(l)} \end{pmatrix}^{\dagger},$$

which shows that $y = \binom{\eta_1}{0} \in \mathcal{R}(A+F^{(l)})$. Consequently, by (2.1) we have $F^{(l)} \in \mathcal{F}_1$ $\forall l$. On the other hand, by (2.5)–(2.6) $\lim_{l\to\infty} F^{(l)} = F \in \mathcal{F}_2$. Hence, we have proved that for any $F \in \mathcal{F}_2$, either $F \in \mathcal{F}_1$ or F is a limit point of a sequence $\{F^{(l)}\}_{l=1}^{\infty}$ in $\mathcal{F}_1$. This means that $\mathcal{F}_2 = \overline{\mathcal{F}}_1$. □

From Lemma 2.2 and (2.4) we get the following corollary immediately.

COROLLARY 2.3. *Let $A, b, y, \mathcal{F}, \mathcal{F}_2$ be as in Lemma 2.2, and let*

$$\hat{A} = A - \begin{pmatrix} \beta_1/\eta_1 & 0 \\ 0 & 0 \end{pmatrix}. \tag{2.7}$$

Then $F \in \mathcal{F}_2$ if and only if $\hat{A} + F$ can be expressed by

$$\hat{A} + F = WC, \tag{2.8}$$

where

$$W = \begin{pmatrix} 0 \\ Q \end{pmatrix} \in \mathcal{R}^{n\times j} \quad \text{with } Q \in \mathcal{R}^{(n-1)\times j} \text{ and } Q^TQ = I, \tag{2.9}$$
$$C \in \mathcal{R}^{j\times m}, \quad 1 \le j \le m-1.$$

It is worth noting two facts: (i) there is not the restriction $\text{rank}(C) = j$ in (2.9), (ii) for any $W \in \mathcal{R}^{n\times j}$ of (2.9), the matrix WC with $C = 0$ corresponds with $F = -\hat{A} \in \mathcal{F}_2$.

3. Optimal bound $\eta(y)$. In this section we study the optimal backward perturbation bound $\eta(y)$ for the coefficient matrix A. The following result presents an explicit expression of the optimal bound.

THEOREM 3.1. *Let $A \in \mathcal{R}^{n\times m}$, nonzero $b \in \mathcal{R}^m$, and nonzero $y \in \mathcal{R}^n$. Let $\eta(y)$ be the optimal backward perturbation bound defined by* (1.2)–(1.3). *Then*

$$\eta(y) = \sqrt{\frac{\|r\|_2^2}{\|y\|_2^2} + \sigma_m^2\left((I - yy^\dagger)A\right)}, \tag{3.1}$$

where $r = b - A^Ty$.

Proof. Expression (3.1) has been proved when $m = 1$ (see (1.5)). We now assume $m > 1$.

By (1.2)–(1.3) and Lemmas 2.1 and 2.2, we have

$$\eta(y) = \inf_{F\in\mathcal{F}} \|F\|_F = \min_{F\in\mathcal{F}_2} \|F\|_F.$$

Expression (3.1) is proved by the following steps.

1. We first consider a special case:

$$y = (\eta_1, 0, \ldots, 0)^T \in \mathcal{R}^n, \quad b = (\beta_1, 0, \ldots, 0)^T \in \mathcal{R}^m, \quad \eta_1, \beta_1 > 0.$$

Write

$$A = (a_1, \ldots, a_m) \quad \text{with} \quad a_i = \begin{pmatrix} \alpha_{i1} \\ a_i^{(0)} \end{pmatrix},\ a_i^{(0)} \in \mathcal{R}^{n-1} \quad \forall i$$

and define the matrices $A^{(0)}$ and $\hat{A}$ by

$$A^{(0)} = (a_1^{(0)}, \ldots, a_m^{(0)}) \in \mathcal{R}^{(n-1)\times m},$$
$$\hat{A} = (\hat{a}_1, a_2, \ldots, a_m) \quad \text{with} \quad \hat{a}_1 = a_1 - (\beta_1/\eta_1, 0, \ldots, 0)^T. \tag{3.2}$$

Then the matrix $\hat{A}$ can also be expressed by

$$\hat{A} = \begin{pmatrix} a^T \\ A^{(0)} \end{pmatrix} \quad \text{with} \quad a = (\alpha_{11} - \beta_1/\eta_1, \alpha_{21}, \ldots, \alpha_{m1})^T. \tag{3.3}$$

1a. By Corollary 2.3, any $F \in \mathcal{F}_2$ can be expressed by $F = WC - \hat{A}$, where W and C are as in (2.9). For an arbitrarily fixed $W = \binom{0}{Q} \in \mathcal{R}^{n\times j}$ $(1 \le j \le m-1)$ of (2.9), we now define a subset $\mathcal{F}_2(W)$ of $\mathcal{F}_2$ by

$$\mathcal{F}_2(W) = \{F = WC - \hat{A} :\ C \in \mathcal{R}^{j\times m}\}.$$

Then by Corollary 2.3

$$\mathcal{F}_2 = \bigcup_W \mathcal{F}_2(W),$$

where $W = \binom{0}{Q} \in \mathcal{R}^{n\times j}$, and Q runs through all $(n-1)\times j$ orthogonal matrices with $j = 1, \ldots, m-1$. Consequently,

$$\begin{aligned}\min_{F\in\mathcal{F}_2} \|F\|_F^2 &= \min_{\substack{W=\binom{0}{Q}\in\mathcal{R}^{n\times j}\\ Q\in\mathcal{R}^{(n-1)\times j}, Q^TQ=I\\ j=1,\ldots,m-1}} \ \min_{F\in\mathcal{F}_2(W)} \|F\|_F^2 \\ &= \min_{\substack{W=\binom{0}{Q}\in\mathcal{R}^{n\times j}\\ Q\in\mathcal{R}^{(n-1)\times j}, Q^TQ=I\\ j=1,\ldots,m-1}} \ \min_{F\in\mathcal{F}_2(W)} \left(\|P_W^{\perp}F\|_F^2 + \|P_W F\|_F^2\right).\end{aligned} \tag{3.4}$$

Let W be an arbitrarily fixed matrix of (2.9), and let $F \in \mathcal{F}_2(W)$. Then by (2.8) we have

$$P_W^{\perp}(\hat{A} + F) = 0, \qquad \|P_W^{\perp}F\|_F = \|P_W^{\perp}\hat{A}\|_F, \tag{3.5}$$

and from the first relation of (3.5)

$$P_W F = \hat{A} + F - P_W\hat{A}. \tag{3.6}$$

We now take $F^* \in \mathcal{R}^{n\times m}$ expressed by

$$F^* = P_W\hat{A} - \hat{A}. \tag{3.7}$$

From

$$\hat{A} + F^* = P_W\hat{A} = WC^* \quad \text{with} \quad C^* = Q^TA^{(0)}$$

we see that $F^* \in \mathcal{F}_2(W)$. Substituting (3.7) into (3.6) gives

$$P_W F^* = 0.$$

Combining it with (3.4) and (3.5), we get

$$\min_{F\in\mathcal{F}_2} \|F\|_F^2 = \min_{\substack{W=\binom{0}{Q}\in\mathcal{R}^{n\times j}\\ Q\in\mathcal{R}^{(n-1)\times j}, Q^TQ=I\\ j=1,\ldots,m-1}} \|P_W^{\perp}\hat{A}\|_F^2, \tag{3.8}$$

where $\hat{A}$ is expressed by (3.3).

1b. By (3.3) and (2.9),

$$\|P_W^{\perp}\hat{A}\|_F^2 = \left\| \begin{pmatrix} 1 & 0 \\ 0 & I - QQ^T \end{pmatrix} \begin{pmatrix} a^T \\ A^{(0)} \end{pmatrix} \right\|_F^2 = \|a\|_2^2 + \|(I - QQ^T)A^{(0)}\|_F^2.$$

Combining it with (3.8) shows that

$$\min_{F\in\mathcal{F}_2} \|F\|_F^2 = \|a\|_2^2 + \|A^{(0)}\|_F^2 - \max_{1\le j\le m-1} \ \max_{\substack{Q\in\mathcal{R}^{(n-1)\times j}\\ Q^TQ=I}} \operatorname{tr}(Q^TA^{(0)}A^{(0)^T}Q), \tag{3.9}$$

where $A^{(0)}$ and a are defined by (3.2) and (3.3), respectively.

Let $\sigma_1(A^{(0)}) \geq \cdots \geq \sigma_m(A^{(0)})$ be the singular values of $A^{(0)}$. Then we have

$$\|A^{(0)}\|_F^2 = \sum_{i=1}^{m} \sigma_i^2(A^{(0)})$$

and [6, p. 191 (4.3.19)]

$$\max_{1\leq j\leq m-1} \max_{\substack{Q\in\mathcal{R}^{(n-1)\times j}\\ Q^TQ=I}} \operatorname{tr}(Q^T A^{(0)} {A^{(0)}}^T Q) = \max_{1\leq j\leq m-1} \sum_{i=1}^{j} \sigma_i^2(A^{(0)}) = \sum_{i=1}^{m-1} \sigma_i^2(A^{(0)}).$$

Moreover, from (3.3)

$$\|a\|_2^2 = \left(\frac{\beta_1}{\eta_1} - \alpha_{11}\right)^2 + \sum_{i=2}^{m} \alpha_{i1}^2.$$

Consequently, (3.9) yields

$$\min_{F\in\mathcal{F}_2} \|F\|_F^2 = \left(\frac{\beta_1}{\eta_1} - \alpha_{11}\right)^2 + \sum_{i=2}^{m} \alpha_{i1}^2 + \sigma_m^2(A^{(0)}). \tag{3.10}$$

2. We now consider the general case: $y \in \mathcal{R}^n, b \in \mathcal{R}^m$, and $y \neq 0$, $b \neq 0$. Let

$$y = U(\tilde{\eta}_1, 0, \ldots, 0)^T = \tilde{\eta}_1 u_1, \quad b = V(\tilde{\beta}_1, 0, \ldots, 0)^T = \tilde{\beta}_1 v_1 \tag{3.11}$$

be the QR factorizations of y and b, respectively, where

$$U = (u_1, U_2) \in \mathcal{R}^{n\times n}, \quad V = (v_1, v_2, \ldots, v_m) \in \mathcal{R}^{m\times m} \tag{3.12}$$

are orthogonal, and $\tilde{\eta}_1, \tilde{\beta}_1 > 0$. Then $(A+F)^T y = b$ can be written as

$$(\tilde{A} + \tilde{F})^T \tilde{y} = \tilde{b},$$

where

$$\tilde{F} = U^T F V, \quad \tilde{A} = U^T A V = \begin{pmatrix} \tilde{\alpha}_{11} & \cdots & \tilde{\alpha}_{m1} \\ & \tilde{A}^{(0)} & \end{pmatrix}$$

with

$$\tilde{\alpha}_{i1} = u_1^T A v_i, \quad i = 1, 2, \ldots, m, \qquad \tilde{A}^{(0)} = U_2^T A V. \tag{3.13}$$

By (3.10),

$$\min_{F\in\mathcal{F}_2} \|F\|_F^2 = \left(\frac{\tilde{\beta}_1}{\tilde{\eta}_1} - \tilde{\alpha}_{11}\right)^2 + \sum_{i=2}^{m} \tilde{\alpha}_{i1}^2 + \sigma_m^2(\tilde{A}^{(0)}), \tag{3.14}$$

where

$$\left(\frac{\tilde{\beta}_1}{\tilde{\eta}_1} - \tilde{\alpha}_{11}\right)^2 + \sum_{i=2}^{m} \tilde{\alpha}_{i1}^2$$

$$= \frac{(v_1^T b - y^T A v_1)^2}{\tilde{\eta}_1^2} + \sum_{i=2}^{m} (u_1^T A v_i)^2 \quad \text{(by (3.11), (3.13))} \tag{3.15}$$

$$= \frac{\|V^T(b - A^T y)\|_2^2}{\|y\|_2^2} = \frac{\|r\|_2^2}{\|y\|_2^2},$$

and

$$\begin{aligned}\sigma_m^2(\tilde{A}^{(0)}) &= \lambda_m(\tilde{A}^{(0)\ T}\tilde{A}^{(0)}) = \lambda_m(A^T U_2 U_2^T A) \quad \text{(by (3.13))}\\ &= \lambda_m(AA^T U_2 U_2^T) = \lambda_m(AA^T(I - u_1 u_1^T)) \quad \text{(by (3.12))}\\ &= \lambda_m(AA^T(I - yy^\dagger)) \quad \text{(by (3.11))}\\ &= \sigma_m^2((I - yy^\dagger)A).\end{aligned}$$

Combining it with (3.14) and (3.15) gives (3.1). □

4. Optimal bound $\eta_\theta(y)$. In this section we apply Theorem 3.1 to derive an explicit expression of $\eta_\theta(y)$ defined by (1.1).

THEOREM 4.1. *Let $A \in \mathcal{R}^{n\times m}, b \in \mathcal{R}^m$, and nonzero $y \in \mathcal{R}^n$, and let $r = b - A^T y$. Define the optimal backward perturbation bound $\eta_\theta(y)$ by* (1.1). *If $b \neq r/(1+\theta^2\|y\|_2^2)$, then*

$$\eta_\theta(y) = \sqrt{\frac{\theta^2\|y\|_2^2}{1+\theta^2\|y\|_2^2}\cdot\frac{\|r\|_2^2}{\|y\|_2^2} + \sigma_m^2\left((I - yy^\dagger)A\right)}. \tag{4.1}$$

Proof. Define the set $\mathcal{H}$ by

$$\mathcal{H} = \left\{ \begin{pmatrix} F \\ g^T \end{pmatrix} \in \mathcal{R}^{(n+1)\times m} : y \text{ is the minimum 2-norm solution to } (A+F)^T y = b + g \right\},$$

and for each fixed $g \in \mathcal{R}^m$, define the vector r_g and the set $\mathcal{H}_g$ by

$$r_g = b + g - A^T y = r + g$$

and

$$\mathcal{H}_g = \{F \in \mathcal{R}^{n\times m} : y \text{ is the minimum 2-norm solution to } (A+F)^T y = b + g\}.$$

Then by (1.1) we have

$$\begin{aligned}[\eta_\theta(y)]^2 &= \inf_{g\in\mathcal{R}^m}\inf_{F\in\mathcal{H}_g}(\|F\|_F^2 + \theta^2\|g\|_2^2)\\ &= \inf_{g\in\mathcal{R}^m}\left(\theta^2\|g\|_2^2 + \inf_{F\in\mathcal{H}_g}\|F\|_F^2\right)\\ &= \inf_{g\in\mathcal{R}^m}\left(\theta^2\|g\|_2^2 + \frac{\|r_g\|_2^2}{\|y\|_2^2} + \sigma_m^2((I - yy^\dagger)A)\right) \quad \text{(by (3.1))}\\ &= \inf_{g\in\mathcal{R}^m}\left(\theta^2\|g\|_2^2 + \frac{\|r+g\|_2^2}{\|y\|_2^2}\right) + \sigma_m^2((I - yy^\dagger)A).\end{aligned} \tag{4.2}$$

Since

$$\theta^2\|g\|_2^2 + \frac{\|r+g\|_2^2}{\|y\|_2^2} = \frac{1+\theta^2\|y\|_2^2}{\|y\|_2^2}\left(\left\|g + \frac{r}{1+\theta^2\|y\|_2^2}\right\|_2^2 + \frac{\theta^2\|y\|_2^2\|r\|_2^2}{(1+\theta^2\|y\|_2^2)^2}\right),$$

we have

$$\inf_{g\in\mathcal{R}^m}\left(\theta^2\|g\|_2^2 + \frac{\|r+g\|_2^2}{\|y\|_2^2}\right) = \frac{\theta^2\|y\|_2^2}{1+\theta^2\|y\|_2^2}\cdot\frac{\|r\|_2^2}{\|y\|_2^2}.$$

Combining it with (4.2) shows (4.1). □

5. Final remarks. In this section we give some remarks on the results of this paper.

Remark 5.1. As Higham [3] pointed out, the formula for $\eta_\theta(y)$ appears to be quite numerically stable. If we are solving the underdetermined system $A^T x = b$ using a QR factorization of A, then we can compute the QR factorization of $A - yy^\dagger A$ in $O(m^2)$ flops using rank-one updating techniques and can then estimate $\sigma_m(A - yy^\dagger A)$ using a condition estimator for triangular matrices; thus the optimal backward perturbation bound (i.e., the backward error) $\eta_\theta(y)$ can be estimated in $O(m^2)$ flops given a QR factorization.

Remark 5.2. Let $\eta_\theta(y)$ and $\eta(y)$ be defined by (1.1) and (1.2)–(1.3), respectively. From (3.1) and (4.1) we see that

$$\eta(y) = \lim_{\theta \to \infty} \eta_\theta(y).$$

Remark 5.3. Let A, b be as in Theorem 3.1, and let nonzero vector y be the minimum 2-norm solution to $A^T x = b$. Then by (1.2) the $n \times m$ matrix $F = 0 \in \mathcal{F}$. By Lemma 2.1, $y \in \mathcal{R}(A)$; i.e., there is a nonzero $z \in \mathcal{R}^m$ such that $y = Az$. Thus we have

$$(I_n - yy^\dagger)Az = (I_n - yy^\dagger)y = 0.$$

This means that $\operatorname{rank}((I_n - yy^\dagger)A) < m$. Consequently,

$$\sigma_m((I_n - yy^\dagger)A) = 0. \tag{5.1}$$

Substituting (5.1) and $r = b - A^T y = 0$ into (3.1) yields $\eta(y) = 0$. From this explanation we can understand why $\eta(y) = 0$ for the exact solution y.

Remark 5.4. If $m \geq n$, then the same argument described in sections 2 and 3 can be used to derive the explicit expression of $\eta(y)$:

$$\eta(y) = \sqrt{\frac{\|r\|_2^2}{\|y\|_2^2} + \sigma_n^2\left((I - yy^\dagger)A\right)} = \frac{\|r\|_2}{\|y\|_2}.$$

Remark 5.5. Let $A \in \mathcal{R}^{n \times m}\,(m \leq n), b \in \mathcal{R}^m$, and nonzero $y \in \mathcal{R}^n$ be given. Define $\eta_*(y)$ by

$$\eta_*(y) = \min\{\|F\|_F : (A + F)^T y = b\}. \tag{5.2}$$

It is well known [8] that

$$\eta_*(y) = \frac{\|r\|_2}{\|y\|_2}, \tag{5.3}$$

where $r = b - A^T y$. Comparing this result with that of Remark 5.4, we see that if $m = n$ and b and y are nonzero vectors, then $\eta(y) = \eta_*(y)$.

Remark 5.6. Let $A \in \mathcal{R}^{n \times m}\,(m \leq n)$, nonzero $b \in \mathcal{R}^m$, and nonzero $y \in \mathcal{R}^n$ be given, and let $\eta(y)$ and $\eta_*(y)$ be defined by (1.2)–(1.3) and (5.2), respectively. Then from expressions (3.1) and (5.3), $\eta(y) \geq \eta_*(y)$. We now give an example where $\eta(y)/\eta_*(y)$ can be arbitrarily large. Let

$$A = \begin{pmatrix} \alpha & 0 \\ 0 & 0 \\ 0 & \alpha \end{pmatrix} \quad \text{with} \quad \alpha = \sqrt{2}(1 - 10^{-k}) \;\; (k \geq 1)$$

and

$$b = (1, 0)^T, \quad y = (1/\sqrt{2}, 1/\sqrt{2}, 0)^T.$$

Then we have

$$r = b - A^T y = \begin{pmatrix} 1 - \alpha/\sqrt{2} \\ 0 \end{pmatrix} = \begin{pmatrix} 10^{-k} \\ 0 \end{pmatrix},$$

$$(I_3 - yy^\dagger)A = \begin{pmatrix} \alpha/2 & 0 \\ -\alpha/2 & 0 \\ 0 & \alpha \end{pmatrix}, \quad \sigma_2((I_3 - yy^\dagger)A) = \alpha/\sqrt{2} = 1 - 10^{-k},$$

$$\eta_*(y) = \frac{\|r\|_2}{\|y\|_2} = 10^{-k},$$

$$\eta(y) = \sqrt{\frac{\|r\|_2^2}{\|y\|_2^2} + \sigma_2^2((I_3 - yy^\dagger)A)} = \sqrt{10^{-2k} + (1 - 10^{-k})^2},$$

and

$$\frac{\eta(y)}{\eta_*(y)} = \sqrt{1 + (10^k - 1)^2} \approx 10^k \gg 1 \ \text{ if } \ k \gg 1.$$

Acknowledgments. We are very grateful to Nick Higham for bringing to our attention the open question of backward error for underdetermined systems and for helpful discussions and valuable suggestions. We are also very grateful to Chris Paige and Urs von Matt, whose pertinent comments inspired us and helped us shorten the analysis and improve the readability of this paper. Chris Paige has developed an alternative proof of Theorem 4.1 based on constructive transformations and reductions.

REFERENCES

[1] R. E. Cline and R. J. Plemmons, *l_2-solutions to underdetermined systems*, SIAM Rev., 18 (1976), pp. 92–106.
[2] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[3] N. J. Higham, private communications, December 1995 and January 1996.
[4] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1996.
[5] D. J. Higham and N. J. Higham, *Backward error and condition of structured linear systems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 162–175.
[6] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.
[7] C. C. Paige, *An error analysis of a method for solving matrix equations*, Math. Comp., 27 (1973), pp. 355–359.
[8] J. L. Rigal and J. Gaches, *On the compatibility of a given solution with the data of a linear system*, J. Assoc. Comput. Mach., 14 (1967), pp. 543–548.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, pp. 403–418, April 1997

011

A FAST PARALLEL CHOLESKY DECOMPOSITION ALGORITHM FOR TRIDIAGONAL SYMMETRIC MATRICES*

ILAN BAR-ON†, BRUNO CODENOTTI‡, AND MAURO LEONCINI§

Abstract. In this paper we present a new parallel algorithm for computing the LL^T decomposition of real symmetric positive-definite tridiagonal matrices. The algorithm consists of a preprocessing and a factoring stage. In the preprocessing stage it determines a rank-$(p-1)$ correction to the original matrix (p = number of processors) by precomputing selected components x_k of the L factor, $k = 1, \ldots, p-1$. In the factoring stage it performs independent factorizations of p matrices of order n/p. The algorithm is especially suited for machines with both vector and processor parallelism, as confirmed by the experiments carried out on a Connection Machine CM5 with 32 nodes. Let $\hat{x}_k$ and $\hat{x}'_k$ denote the components computed in the preprocessing stage and the corresponding values (re)computed in the factorization stage, respectively. Assuming that $|\hat{x}_k/\hat{x}'_k|$ is small, $k = 1, \ldots, p-1$, we are able to prove that the algorithm is stable in the backward sense. The above assumption is justified both experimentally and theoretically. In fact, we have found experimentally that $|\hat{x}_k/\hat{x}'_k|$ is small even for ill-conditioned matrices, and we have proven by an a priori analysis that the above ratios are small provided that preprocessing is performed with suitably larger precision.

Key words. parallel algorithm; Cholesky decomposition; LR and QR algorithms; eigenvalues; symmetric, tridiagonal, and band matrices; CM5

AMS subject classifications. 15A18, 15A23, 65F05, 65F15, 65Y05

PII. S0895479895282623

1. Introduction.

We consider the problem of computing the Cholesky decomposition of very large real symmetric positive-definite tridiagonal matrices. Cholesky decomposition is a valuable tool in many diagonalization techniques for computing eigenvalues and singular values of matrices. Rutishauser's cubically convergent LR algorithm is based on the iterative application of Cholesky decomposition [21]. The divide-and-conquer approach can also be combined with it [4, 5]. More recently, the Cholesky decomposition, or one of its variants, has been used in connection with the accurate computation of the singular values of bidiagonal matrices [11, 15] and the eigenvalues of specially structured symmetric tridiagonal matrices [9]. Moreover, it has been shown that Francis's QR algorithm (see [16, 17]) can be implemented using a band Cholesky decomposition [3].

Cholesky decomposition, followed by the parallel solution of the respective bidiagonal systems [8], is one of the most natural approaches to the solution of positive-definite linear systems [2, 12, 23, 24, 26], but as such has not received a great deal of attention.

The classical sequential algorithms for computing the Cholesky decomposition

* Received by the editors February 24, 1995; accepted for publication (in revised form) by K. Sigmon May 27, 1996. This research was produced with the help of the National Science Foundation, Infrastructure grant CDA-8722788, and Cooperation Agreement CNR-MOSA.

http://www.siam.org/journals/simax/18-2/28262.html

† Department of Computer Science, Technion, Haifa 32000, Israel. Present address: Science and Mathematics, University of Texas of the Permian Basin, 4901 E. University, Odessa, TX 79762 (baron_i@utpb.edu).

‡ Istituto di Matematica Computazionale del CNR, Via S. Maria 46, 56126 Pisa, Italy. This research was supported by the ESPRIT III Basic Research Programme of the EC under contract 9072 (Project GEPPCOM).

§ Dipartimento di Informatica, Università di Pisa, Corso Italia 40, 56125 Pisa, Italy (leoncini@di.unipi.it). This research was supported by the ESPRIT III Basic Research Programme of the EC under contract 9072 (Project GEPPCOM) and by M.U.R.S.T. 40% funds.

cannot be efficiently parallelized or directly vectorized. It is thus natural to seek an algorithm directly amenable for an efficient parallel implementation. In this paper we introduce a new algorithm which borrows ideas from the substructured parallel cyclic reduction algorithm for the solution of tridiagonal systems [19, 28]. Parallel cyclic reduction consists of three stages.

1. (Almost) local forward and backward Gaussian elimination steps. During this stage only one communication is required, usually with an adjacent processor.
2. Solution of a reduced system with one equation per processor.
3. Local backsubstitution.

Our algorithm consists of three stages as well. Let A be an $N \times N$ tridiagonal matrix and assume for simplicity that $N = np$, where p is the number of available processors. Also, viewing A as block partitioned, let T_i denote its $n \times n$ diagonal blocks, $i = 1, \ldots, p$. Finally, let L be the Cholesky factor of A, and let x_i, $i = 1, \ldots, N$, denote its diagonal elements. The stages are as follows.

1. Local forward and backward Gaussian elimination steps. This stage, which requires no communication, returns a reduced tridiagonal matrix B of order $2p - 3$.
2. Computation of $x_{n(i-1)}$, $i = 2, \ldots, p$, by applying suitable transformations to B. This is the only stage which requires (*tree-like* structured) communications between the processors.
3. Local factorization of p $n \times n$ matrices T_i', where $T_1' = T_1$ and T_i', $i = 2, \ldots, p$, is a rank-one update (involving $x_{n(i-1)}$) of T_i.

We refer to 1 and 2 together as the *preprocessing* stage.

The time complexity of our algorithm is about $8\frac{N}{p} + 15 \log p$ if p processors are available. If $\log p \ll \frac{N}{p}$ the complexity is governed by the factor $8n$. Under this circumstance, the parallel algorithm requires about four times the number of flops of the classical sequential algorithm, with a (theoretical) speedup close to $p/4$. We show that our parallel algorithm is also computationally efficient in practice. We report the results obtained on a Connection Machine CM5 supercomputer with 32 nodes and 128 vector units altogether [27]. We achieve very satisfactory performances on large matrices (say $N \geq 2^{18}$). For smaller size matrices, very good performances can still be obtained by appropriately scaling down the number of processors involved.

A natural competitor with our algorithm is the recursive doubling algorithm for the LU decomposition of tridiagonal matrices [24]. When recursive doubling is used in the LR algorithm (and to compute Cholesky rather than LU decomposition) it achieves parallel complexity roughly $\sim 12 \log N$ using an unbounded (i.e., linear in N) number of processors [25]. In the more realistic case of $p \ll N$, and using parallel prefix instead of recursive doubling, the parallel time complexity becomes roughly $\sim 27\frac{N}{p}$, which is more than three times larger than ours.

Cholesky decomposition is componentwise stable, and the variant presented here retains this property. With respect to the classical algorithm, the backward error affecting the coefficient matrix is further influenced, in the first diagonal entry of each block T_i, by the factor $|\hat{x}_{n(i-1)}/\hat{x}'_{n(i-1)}|$, $i = 2, \ldots, p$. Here $\hat{x}_{n(i-1)}$ and $\hat{x}'_{n(i-1)}$ denote the components of the L factor computed in the preprocessing and recomputed in the actual factorization stage, respectively. We find experimentally that these ratios are small even for ill-conditioned matrices. We have also proven, by an a priori analysis, that $|\hat{x}_{n(i-1)}/\hat{x}'_{n(i-1)}|$ are small provided that preprocessing is performed with suitably larger precision.

This paper is organized as follows. In section 2 we define concepts and notation

used throughout the rest of the paper. In section 3 we review the LR algorithm for computing the eigenvalues of symmetric tridiagonal matrices. This will provide motivation for the development of a parallel algorithm for computing the Cholesky decomposition of such matrices. In section 4 we describe a sequential algorithm that computes the Cholesky factors and discuss its implementation, computational cost, and numerical accuracy. In section 5 we describe the parallel algorithm, providing details of the preprocessing stage, and, in section 6, we analyze its computational cost and suitability to vectorization. In section 7 we present the experimental results obtained on the CM5, and in section 8 we present the error analysis which shows the numerical accuracy of the algorithm. We conclude with some suggestions for further work.

2. Definitions and main notation. We denote by $\mathcal{R}^n$ the set of real vectors of order n and by e_i the n-vector whose entries are all zero except the ith one, which is 1. When needed, we emphasize that a particular vector e_i is in $\mathcal{R}^n$ by writing $e_i^{(n)}$.

We denote by $\mathcal{M}(n)$ the set of real $n \times n$ matrices and by A^T the transpose of A.

We denote a tridiagonal symmetric matrix $T \in \mathcal{M}(n)$ by

$$T = \begin{pmatrix} a_1 & b_2 & & & \\ b_2 & a_2 & b_3 & & \\ & b_3 & & \ddots & \\ & & \ddots & & b_n \\ & & & b_n & a_n \end{pmatrix}. \tag{1}$$

In this paper we assume that T is unreduced, that is, $b_i \neq 0,\ i = 2, \ldots, n$.

We say that a nonsingular matrix $P \in \mathcal{M}(m)$ is a *cyclic transformation* if

$$P = \begin{pmatrix} H & 0 \\ h^T & 1 \end{pmatrix}, \qquad \begin{matrix} H \in \mathcal{M}(m-1), \\ h \in \mathcal{R}^{(m-1)}. \end{matrix}$$

Note that by this definition H is nonsingular as well.

We say that the computation of the Cholesky decomposition of a matrix A is *componentwise stable* if the computed Cholesky factors are the exact decomposition of a small componentwise perturbation of A.

We measure the time complexity of a sequential algorithm by counting the number of *flops*, i.e., floating point operations. We also refer to the flop count as the number of (arithmetic) *steps*. The time complexity of a parallel algorithm implemented on a p processor machine is the maximum, over the p processors, of the number of steps performed. We refer to this measure as the number of *parallel steps*.

The *speedup* of a parallel algorithm A over a sequential algorithm B is the ratio

$$S_p(n) = \frac{T_B(n)}{T_{A,p}(n)},$$

where $T_B(n)$ is the (time) complexity of B on inputs of size n and $T_{A,p}(n)$ is the complexity of A on inputs of size n with p processors. Obviously, for any parallel algorithm there is some sequential algorithm for which $S_p(n) \leq p$, for otherwise a sequential simulation of the parallel algorithm would beat the (supposedly) best known sequential one. However, in this paper we are interested in comparing the running time of the parallel algorithm with that of the classical sequential method. Hence, we may obtain superlinear speedups due to a more efficient use of the architecture resources, namely, data transmission and vectorization.

3. An overview of the LR algorithm. The LR algorithm developed by Rutishauser was termed by Wilkinson "the most significant advance which has been made in connection with the eigenvalue problem since the advent of automatic computers" (see [29, p. 485]). This algorithm is very simple and efficient and computes the eigenvalues of tridiagonal symmetric matrices with a cubic rate of convergence.

The LR algorithm iteratively computes a sequence of tridiagonal matrices that gradually converge to a diagonal matrix with the same eigenvalues. Starting with the original matrix $A_0 = A$ and with $eig = 0$, for $s = 0, 1, \ldots$, the sth step consists of the following stages:

- choose an appropriate shift y_s,
- find the Cholesky decomposition of $B_s = A_s - y_s I = L_s L_s^T$,
- set $A_{s+1} = L_s^T L_s$ and $eig = eig + y_s$.

As soon as the last off-diagonal element becomes negligible, eig is a new exposed eigenvalue. It is easy to see that the third stage of this algorithm can be efficiently parallelized. In addition, after a few steps, the shifts y_s in the first stage can be read off the last diagonal element of the matrix (see Rutishauser and Schwarz [22]). It follows that the main difficulty in implementing the LR algorithm on a parallel machine lies in the Cholesky decomposition. This is one major motivation to focus our attention on the development of an efficient parallel implementation of Cholesky decomposition. For further discussions on the LR algorithm the reader is encouraged to see Wilkinson [29], Parlett [20], Grad and Zakrajšek [18], and Bar-On [3].

4. Cholesky decomposition. In this section we describe a sequential algorithm to compute the Cholesky decomposition of a symmetric tridiagonal matrix which is particularly suitable to implement the LR algorithm and analyze its computational and numerical properties.

Consider the Cholesky decomposition stage in the LR algorithm described in section 3 and let (1) be the matrix to be factored. We have that

$$T = \begin{pmatrix} d_1 & & & & \\ y_2 & d_2 & & & \\ & y_3 & d_3 & & \\ & & \ddots & \ddots & \\ & & & y_n & d_n \end{pmatrix} \begin{pmatrix} d_1 & y_2 & & & \\ & d_2 & y_3 & & \\ & & \ddots & \ddots & \\ & & & d_{n-1} & y_n \\ & & & & d_n \end{pmatrix} = LL^T. \tag{2}$$

Instead of computing decomposition (2) and taking into account that this process must be repeatedly applied over LR iterations, we compute the quantities x_i and z_i using the following recurrences:

$$z_i = b_i^2 / x_{i-1}, \quad x_i = a_i - z_i, \quad i = 1, \ldots, n, \tag{3}$$

with $x_0 = 1$. Note that in recurrences (3) we use only the a_i's and b_i^2's (rather than the b_i's). It can be easily proved by induction that $x_i = d_i^2$ and $z_i = y_i^2$. Now, if we set

$$L^T L = \begin{pmatrix} f_1 & g_2 & & & \\ g_2 & f_2 & g_3 & & \\ & g_3 & & \ddots & \\ & & \ddots & & g_n \\ & & & g_n & f_n \end{pmatrix},$$

then we can efficiently compute the quantities f_i and g_i^2 as follows:

$$g_i^2 = z_i * x_i, \quad f_i = x_i + z_{i+1}, \quad i = 1, \ldots, n,$$

with $z_{n+1} = 0$.

This process can therefore be iterated. If needed, the elements of the matrix (implicitly) generated at the ith step of the LR algorithm can be easily recovered. By using this variant of the Cholesky decomposition, which we call the *revised* decomposition, we avoid the computation of square roots.

Complexity. The purpose of this paragraph is to point out the rather poor performance that one gets by using both classical and revised Cholesky decompositions on sequential computers. Table 1 shows the running times, observed on a DEC Alpha 7000 Model 660 Super Scalar machine, of the following routines: the BLAS routine "dgemm" which performs matrix multiplication; the LAPACK routines "dpotrf" and "dpbtrf" [1] which perform the Cholesky decomposition on dense and tridiagonal matrices, respectively; and the private routine "trid" which computes the above revised decomposition. The revised decomposition is more efficient than the classical one primarily because of the absence of square root computations. However, the Mflops column shows that it is still very inefficient with respect to the dense computations dgemm and dpotrf mainly because of the low number of flops per memory reference.

TABLE 1
LAPACK computational routines.

Routine	n	Flops	Time	Mflops
dgemm	400	$2 * n^3$	0.95	135.48
dpotrf	600	$2 * n^3/6$	0.99	72.11
dpbtrf	200000	$2 * n$	1.01	0.39
trid	200000	$2 * n$	0.08	5.00

Numerical stability. Cholesky decomposition is componentwise stable, and this variant retains this property. Usually the entries of the given matrix are known up to some perturbation so that it is very useful to investigate the "structure" of the perturbations introduced by rounding. To show this, let us denote the computed value of a by $\hat{a} = fl(a)$ and assume that the standard operations satisfy

$$fl(a \; op \; b) = (a \; op \; b)(1 + \eta), \quad |\eta| \leq \theta,$$

where op stands for $+, -, *$, or $/$ and θ is the machine relative precision. For example, θ is roughly 10^{-16} in standard double precision. Then the actual computation of the decomposition can be formulated as follows:

$$\begin{array}{l} \hat{z}_i = c_i(1 + \beta_i)/\hat{x}_{i-1} = \hat{c}_i/\hat{x}_{i-1}, \\ \hat{x}_i = (a_i - \hat{z}_i)(1 + \alpha_i') = \hat{a}_i - \hat{z}_i, \end{array} \qquad i = 1, \ldots, n,$$

with $c_i = b_i^2$, $\hat{a}_i = a_i(1 + \alpha_i)$, and $|\beta_i| \leq \theta$, $|\alpha_i| \leq |\alpha_i'| \leq \theta$. For the classical error bounds for Cholesky decomposition see [30] and [14].

5. Parallel Cholesky decomposition.

5.1. Mathematical formulation. Let $T \in \mathcal{M}(n)$ be the unreduced symmetric tridiagonal matrix in (1). In block notation, T can be written as

$$T = \begin{pmatrix} T_1 & U_2^T & & & \\ U_2 & T_2 & U_3^T & & \\ & U_3 & \ddots & \ddots & \\ & & \ddots & \ddots & U_q^T \\ & & & U_q & T_q \end{pmatrix}, \qquad \begin{gathered} T_i \in \mathcal{M}(n_i), \\ \sum_{i=1}^{q} n_i = n, \\ m_i = \sum_{j=1}^{i} n_j, \\ U_{i+1} = b_{m_i+1} e_1^{(n_{i+1})} \left(e_{n_i}^{(n_i)}\right)^T \end{gathered}$$

and the Cholesky factors of the decomposition in (2) as

$$L = \begin{pmatrix} L_1 & & & & \\ R_2 & L_2 & & & \\ & R_3 & L_3 & & \\ & & \ddots & \ddots & \\ & & & R_q & L_q \end{pmatrix}, \qquad \begin{gathered} L_i \in \mathcal{M}(n_i), \\ R_{i+1} = y_{m_i+1} e_1^{(n_{i+1})} \left(e_{n_i}^{(n_i)}\right)^T. \end{gathered}$$

By equating LL^T and T we obtain

$$T_1' \equiv L_1 L_1^T = T_1$$

and

$$T_k' \equiv L_k L_k^T = T_k - R_k R_k^T = T_k - y_{m_{k-1}+1}^2 e_1^{(n_k)} (e_1^{(n_k)})^T, \qquad k = 2, \ldots, q.$$

Our parallel algorithm precomputes the "perturbations" $y_{m_{k-1}+1}^2$ and then applies the transformation

$$a_{m_{k-1}+1}' \equiv a_{m_{k-1}+1} - y_{m_{k-1}+1}^2, \qquad k = 2, \ldots, q,$$

thus reducing the computation of the Cholesky decomposition of T to q independent instances of the same problem, i.e., the computation of the Cholesky factors of $T_1', \ldots, T_q'$. We now show some preliminary facts about these perturbations that we will later use to prove the correctness of our parallel algorithm.

For $i = 1, \ldots, n$, let

$$T_{(i)} = \begin{pmatrix} a_1 & b_2 & & & \\ b_2 & a_2 & b_3 & & \\ & b_3 & & \ddots & \\ & & \ddots & & b_i \\ & & & b_i & a_i \end{pmatrix}.$$

Then it easily follows that

$$y_{m_k+1}^2 = b_{m_k+1}^2 e_{m_k}^T T_{(m_k)}^{-1} e_{m_k}.$$

Actually, our parallel algorithm does not explicitly compute the perturbation $y_{m_k+1}^2$ of the first diagonal element of the block T_{k+1}. Instead, it computes the quantity

$$x_{m_k} \equiv a_{m_k}' = a_{m_k} - b_{m_k}^2 e_{m_k-1}^T T_{(m_k-1)}^{-1} e_{m_k-1} \tag{4}$$

(i.e., a perturbed last element of the preceding block) and then obtains

$$a'_{m_k+1} = a_{m_k+1} - b^2_{m_k+1}/x_{m_k}$$

by using recurrences (3). The perturbation originally sought can be expressed, in terms of the computed quantity, as $y^2_{m_k+1} = b^2_{m_k+1}/x_{m_k}$.

LEMMA 5.1. *Let $P_i, i = 1, \ldots, j$, be a sequence of cyclic transformations. Then*

$$P = P_j \cdots P_2 P_1 = \prod_{i=1}^{j} \begin{pmatrix} H_i & 0 \\ h_i^T & 1 \end{pmatrix} = \begin{pmatrix} H & 0 \\ h^T & 1 \end{pmatrix}$$

is a cyclic transformation.

LEMMA 5.2. *Let $P_{(m_k)}$ be a cyclic transformation such that*

$$(5)\; P_{(m_k)} T_{(m_k)} = \begin{pmatrix} H_k & 0 \\ h_k^T & 1 \end{pmatrix} \begin{pmatrix} T_{(m_k-1)} & b_{m_k} e_{m_k-1} \\ b_{m_k} e^T_{m_k-1} & a_{m_k} \end{pmatrix} = \begin{pmatrix} \tilde{T}_{(m_k-1)} & * \\ 0 & \tilde{a}_{m_k} \end{pmatrix}.$$

Then we have

$$\tilde{a}_{m_k} = a_{m_k} + b_{m_k} h_k^T e_{m_k-1} = x_{m_k}. \tag{6}$$

Proof. From the second equality in (5) we have $T_{(m_k-1)} h_k = -b_{m_k} e_{m_k-1}$; hence $h_k = -b_{m_k} T^{-1}_{(m_k-1)} e_{m_k-1}$ and (6) follows from (4). ∎

In the preprocessing stage of our parallel algorithm we apply a sequence of parallel cyclic transformations to obtain the values x_{m_k}, $k = 1, \ldots, q-1$, called *pivots*.

5.2. The algorithm. We assume for simplicity that the tridiagonal matrix is of order $N = np$, with p being the number of processors. We initially distribute the entries of the matrix between the processors so that each processor stores n consecutive rows. We denote these blocks of rows by

$$B_i = \begin{pmatrix} b_{(i-1)n+1} & a_{(i-1)n+1} & b_{(i-1)n+2} & & \\ \ddots & \ddots & \ddots & & \\ & b_{in} & a_{in} & b_{in+1} & \end{pmatrix} \in \mathcal{M}(n, n+2)$$

for $i = 1, \ldots, p$.

Our *parallel Cholesky* algorithm consists of three stages:

(i) Diagonalization,

(ii) Bottom-Up and Top-Down sweeps,

(iii) Factorization.

In stage (i) each processor performs locally $O(n)$ parallel steps independently. In stage (ii) the processors perform $O(\log p)$ operations which require interprocessor communication. Finally, in stage (iii) each processor performs $O(n)$ parallel steps independently. Altogether, the number of parallel steps is $O(n + \log p)$.

Stage (i): *Diagonalization.* Let

$$B = \begin{pmatrix} b_1 & a_1 & z^T & & \\ & z & A & c & \\ & & c^T & a_n & b_{n+1} \end{pmatrix} \in \mathcal{M}(n, n+2) \tag{7}$$

denote the block assigned to a generic processor, where A is a tridiagonal matrix of order $n-2$, $z = b_2 e_1^{(n-2)}$, and $c = b_n e_{n-2}^{(n-2)}$. Each processor i, $1 < i < p$, performs

a forward Gaussian elimination procedure to eliminate b_n in the last row and then a backward Gaussian elimination procedure to eliminate b_2 in the first row. In matrix notation this amounts to applying a cyclic transformation

$$B' = PB = \begin{pmatrix} 1 & -b_2 f^T & \\ & I & \\ & -b_n g^T & 1 \end{pmatrix} B, \qquad \begin{array}{l} f = A^{-1} e_1^{(n-2)}, \\ g = A^{-1} e_{n-2}^{(n-2)}, \end{array}$$

so that

$$B' = \begin{pmatrix} b_1 & a_1' & & b_2' & \\ & z & A & c & \\ & b_n' & & a_n' & b_{n+1} \end{pmatrix} = \begin{pmatrix} b_1 & v & & y & \\ & z & A & c & \\ & y & & w & b_{n+1} \end{pmatrix}, \tag{8}$$

since $y = b_2' = b_n'$ by symmetry. Processor 1 performs the forward Gaussian elimination step only to eliminate b_n in the last row, obtaining

$$B' = \begin{pmatrix} A & c & \\ & w & b_{n+1} \end{pmatrix} \in \mathcal{M}(n, n+1),$$

and processor p remains idle. By the end of stage (i), the in-between rows and columns do not further contribute to the search for the pivots x_{m_k}, and they can be ignored. We now consider the matrix $T^{(0)}$, of order $2p-3$, formed using the relevant elements from the blocks B' computed by processors 1 through $p-1$. Note that processor 1 contributes one row, while processor i, $1 < i < p$, contributes two rows; i.e.,

$$T^{(0)} \equiv \begin{pmatrix} w_1^{(0)} & b_2^{(0)} & & & \\ b_2^{(0)} & & \ddots & & \\ & \ddots & \ddots & b_{p-1}^{(0)} & \\ & & b_{p-1}^{(0)} & v_{p-1}^{(0)} & y_{p-1}^{(0)} \\ & & & y_{p-1}^{(0)} & w_{p-1}^{(0)} \end{pmatrix},$$

where $b_{i+1}^{(0)} = b_{in+1}$. Processor $i = 2, \ldots, p-1$ stores the submatrix

$$T_i^{(0)} \equiv \begin{pmatrix} b_i^{(0)} & v_i^{(0)} & y_i^{(0)} & \\ & y_i^{(0)} & w_i^{(0)} & b_{i+1}^{(0)} \end{pmatrix},$$

while processor 1 stores

$$X_1^{(0)} \equiv \begin{pmatrix} w_1^{(0)} & b_2^{(0)} \end{pmatrix} \equiv \begin{pmatrix} x_1 & b_2^{(0)} \end{pmatrix}.$$

Stage (ii): *Bottom-Up and Top-Down sweeps.* This stage consists of two sequences of cyclic transformations, called Bottom-Up and Top-Down sweeps, which involve the submatrices stored in the different processors according to a tree-like pattern. Each *sweep* corresponds to the *merging* of two submatrices with the generation of a new submatrix using the extreme rows.

Bottom-Up sweeps are performed as follows. For $s = 1, \ldots, \log_2 p - 1$ and $i = 2, \ldots, p/2^s - 1$, first merge the matrices $T_{2i-1}^{(s-1)}$ and $T_{2i}^{(s-1)}$,

$$\begin{pmatrix} T_{2i-1}^{(s-1)} \\ T_{2i}^{(s-1)} \end{pmatrix} = \begin{pmatrix} b_{2i-1}^{(s-1)} & v_{2i-1}^{(s-1)} & y_{2i-1}^{(s-1)} & & & \\ & y_{2i-1}^{(s-1)} & w_{2i-1}^{(s-1)} & b_{2i}^{(s-1)} & & \\ & & b_{2i}^{(s-1)} & v_{2i}^{(s-1)} & y_{2i}^{(s-1)} & \\ & & & y_{2i}^{(s-1)} & w_{2i}^{(s-1)} & b_{2i+1}^{(s-1)} \end{pmatrix},$$

and then eliminate $y_{2i-1}^{(s-1)}$ in the top row and $y_{2i}^{(s-1)}$ in the bottom row by applying a cyclic transformation $P_i^{(s)}$; i.e.,

$$P_i^{(s)} \begin{pmatrix} T_{2i-1}^{(s-1)} \\ T_{2i}^{(s-1)} \end{pmatrix} = \begin{pmatrix} b_i^{(s)} & v_i^{(s)} & & & y_i^{(s)} & \\ & y_{2i-1}^{(s-1)} & w_{2i-1}^{(s-1)} & b_{2i}^{(s-1)} & & \\ & & b_{2i}^{(s-1)} & v_{2i}^{(s-1)} & y_{2i}^{(s-1)} & \\ & y_i^{(s)} & & & w_i^{(s)} & b_{i+1}^{(s)} \end{pmatrix}.$$

Finally, form the matrix $T_i^{(s)}$ using the extreme rows

$$T_i^{(s)} = \begin{pmatrix} b_i^{(s)} & v_i^{(s)} & y_i^{(s)} & \\ & y_i^{(s)} & w_i^{(s)} & b_{i+1}^{(s)} \end{pmatrix}.$$

For $i = 1$ the merging operation involves only three rows,

$$\begin{pmatrix} X_1^{(s-1)} \\ T_2^{(s-1)} \end{pmatrix} = \begin{pmatrix} x_{2^{s-1}} & b_2^{(s-1)} & & \\ b_2^{(s-1)} & v_2^{(s-1)} & y_2^{(s-1)} & \\ & y_2^{(s-1)} & w_2^{(s-1)} & b_3^{(s-1)} \end{pmatrix},$$

and $y_2^{(s-1)}$ is eliminated from the bottom row

$$P_1^{(s)} \begin{pmatrix} X_1^{(s-1)} \\ T_2^{(s-1)} \end{pmatrix} = \begin{pmatrix} x_{2^{s-1}} & b_2^{(s-1)} & & \\ b_2^{(s-1)} & v_2^{(s-1)} & y_2^{(s-1)} & \\ & & w_1^{(s)} & b_2^{(s)} \end{pmatrix}, \tag{9}$$

yielding the matrix

$$X_1^{(s)} = \begin{pmatrix} w_1^{(s)} & b_2^{(s)} \end{pmatrix} \equiv \begin{pmatrix} x_{2^s} & b_2^{(s)} \end{pmatrix}.$$

The Top-Down sweeps are performed in a similar way.

For $s = \log_2 p - 2, \log_2 p - 3, \ldots, 0$, and odd i (i.e., $i = 3, 5, \ldots, p/2^s - 1$), let the nonnegative integer l and the positive odd j be such that $i = 2^l j + 1$ (l and j are uniquely determined). First merge $X_j^{(s+l)}$ and $T_i^{(s)}$:

$$\begin{pmatrix} X_j^{(s+l)} \\ T_i^{(s)} \end{pmatrix} = \begin{pmatrix} x_{j2^{s+l}} & b_i^{(s)} & & \\ b_i^{(s)} & v_i^{(s)} & y_i^{(s)} & \\ & y_i^{(s)} & w_i^{(s)} & b_{i+1}^{(s)} \end{pmatrix}.$$

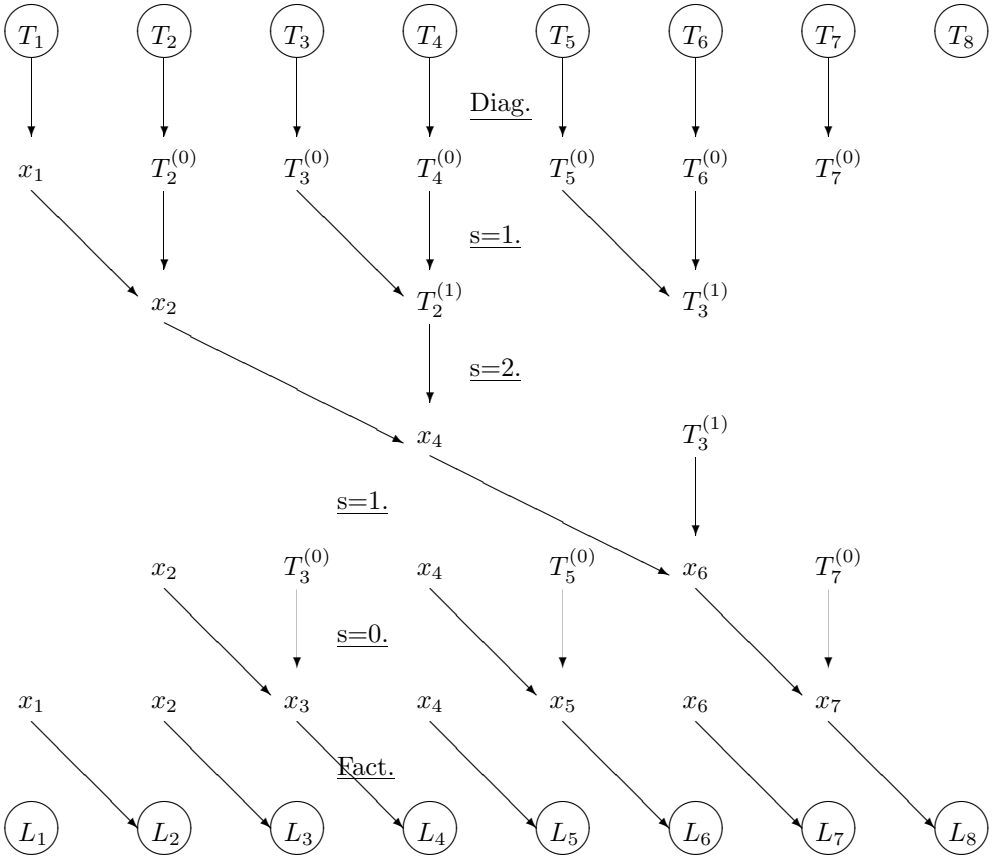


FIG. 1. *A flowchart for $p = 2^3$ processors.*

Then eliminate $y_i^{(s)}$ in the bottom row and form $X_i^{(s)} = (x_{i2^s} \quad b_{i+1}^{(s)})$.

Bottom-Up and Top-Down sweeps for $p = 8$ are depicted in Figure 1.

THEOREM 5.3. *Stages* (i) *and* (ii) *of the parallel Cholesky algorithm correctly compute the pivots $x_k, k = 1, \ldots, p-1$.*

Proof. With respect to any processor $k, 1 \leq k \leq (p-1)$, each transformation computed during stages (i) and (ii) of the algorithm is a cyclic transformation applied to the submatrix $T_{(m_k)}$. It follows from Lemma 5.1 that the whole sequence is still a cyclic transformation applied to $T_{(m_k)}$. Since this annihilates the off-diagonal element b_{m_k}, the proof follows from Lemma 5.2. ∎

Note that the pivotal elements x_{2^s} are computed from the tridiagonal symmetric

matrix

$$\mathcal{T}_{2^s} = \begin{pmatrix} X_1^{(s-1)} & & & & & \\ & T_3^{(s-2)} & & & & \\ & & \ddots & & & \\ & & & T_{2^s-1}^{(0)} & & \\ & & & & T_{2^s} \end{pmatrix}$$

of order $n+2s-1$. Analogously, for $i = 2^l j+1$, the pivot elements x_{i2^s} are computed from the tridiagonal symmetric matrix

$$\mathcal{T}_{i2^s} = \begin{pmatrix} X_j^{(s+l)} & & & & & & \\ & T_{2i-1}^{(s-1)} & & & & & \\ & & T_{4i-1}^{(s-2)} & & & & \\ & & & \ddots & & & \\ & & & & T_{i2^s-1}^{(0)} & & \\ & & & & & T_{i2^s} \end{pmatrix}$$

of order $n+2s+1$.

Stage (iii): *Factorization.* The parallel factorization of the independent blocks is straightforward. Processor 1 computes the Cholesky decomposition of its original block T_1, while processors $i = 2, \ldots, p$ modify their blocks according to the rule

$$a'_{(i-1)n+1} = a_{(i-1)n+1} - b^2_{(i-1)n+1}/x_{(i-1)n}$$

and then compute their decompositions.

6. Parallel computational cost. In this section we study the computational cost of the parallel Cholesky algorithm of section 5.2.

Stage (i). To determine the cost of this stage, we must give the details of the forward and backward Gaussian elimination procedures. We denote the blocks in each processor as in (7) and the computed transformations as in (8). Since we compute the revised decomposition introduced in section 4, in what follows we actually use the squares c_i of the off-diagonal elements b_i of the matrix T.

• Forward Gaussian elimination:

1. Set $z = c_2$ and $w = a_2$.
2. For $i = 3, \ldots, n$, set

$$\begin{aligned} t &= c_i/w, \\ z &= z * t/w, \\ w &= a_i - t. \end{aligned}$$

• Backward Gaussian elimination:

1. Set $v = a_{n-1}$.
2. For $i = n-1, \ldots, 2$, set

$$\begin{aligned} t &= c_i/v, \\ v &= a_{i-1} - t. \end{aligned}$$

The flop count for stage (i) is therefore $\sim 6n$.

Stage (ii). Let

$$T = \begin{pmatrix} b_1 & v_1 & y_1 & & & \\ & y_1 & w_1 & b_2 & & \\ & & b_2 & v_2 & y_2 & \\ & & & y_2 & w_2 & b_3 \end{pmatrix} \Rightarrow \begin{pmatrix} b_1 & v & & & y & \\ & y_1 & w_1 & b_2 & & \\ & & b_2 & v_2 & y_2 & \\ & y & & & w & b_3 \end{pmatrix}$$

denote a typical transformation in the Bottom-Up sweep. Again, we consider the squares of the off-diagonal elements $c_2 = b_2^2$ and $z_i = y_i^2$, $i = 1, 2$, and compute $z = y^2$. Therefore, we perform the following calculation:

$$\begin{gathered} \alpha = w_1 * v_2, \qquad \beta = c_2/\alpha, \qquad \gamma = 1 - \beta, \\ t_1 = z_1/(w_1 * \gamma), \qquad t_2 = z_2/(v_2 * \gamma), \\ v = v_1 - t_1, \qquad w = w_2 - t_2, \qquad z = \beta * t_1 * t_2, \end{gathered}$$

which takes 11 parallel steps.

Similarly, let

$$T = \begin{pmatrix} x_1 & b_2 & & \\ b_2 & v_2 & y_2 & \\ & y_2 & w_2 & b_3 \end{pmatrix} \Rightarrow \begin{pmatrix} x_1 & b_2 & & \\ b_2 & v_2 & y_2 & \\ & & x_2 & b_3 \end{pmatrix}$$

denote a typical transformation in the Top-Down sweep. Then we compute $t = v_2 - c_2/x_1$ and $x_2 = w_2 - z_2/t$ in four parallel steps.

Stage (iii). The number of parallel steps is $\sim 2n$ (see section 4).

The total number T_p of parallel steps is therefore

$$T_p \sim 6n + 11 \log p + 4 \log p + 2n = 8n + 15 \log p.$$

Assuming that $\log p \ll n$, we conclude that the cost of the parallel algorithm is governed by the factor $8n$. Hence, the parallel algorithm requires about four times the number of flops of the sequential algorithm. The theoretical speedup is thus $p/4$. However, on vector, pipelined, and super-scalar machines, the flop count determines the true performance of an algorithm only to within a constant factor. Actually, an algorithm with a worse flop count might perform better in the case it can be vectorized. We show now that our parallel algorithm can be satisfactorily vectorized.

Vectorization. Let $N = pn$, where $p = 2^r$ is the number of "physical" processors, each with vectorization capability. One possibility for exploiting this additional power relies on employing some *parallel slackness*. In other words, we assume that the number of available processors is larger than p and let each physical processor simulate many such "logical" processors. More precisely, let $n = mP$, where $P = 2^t$, so that $N = qm$, with $q = pP = 2^{r+t}$. We let each physical processor perform the tasks of P corresponding logical processors. The number of flops in stages (i) and (iii) is still approximately $8n$. The number of flops in stage (ii) increases to about $15(r + P)$, which is still negligible for $(r + P) \ll n$. However, the main stages of the algorithm, namely, stages (i) and (iii), can now be vectorized, with each processor working on vectors of length $P = 2^t$. We provide an example of this sort in the next section.

7. Numerical examples. In this section we present some experimental results obtained on a CM5 parallel supercomputer with $p = 32$ nodes. Each node is in turn composed of four vector units, controlled by a SPARC microprocessor, and 32 Mbytes of memory. The running time and speedup for the largest problems on which we could experiment are shown in Table 2. We have computed the Cholesky factorization of several classes of tridiagonal matrices, including (a) the symmetric tridiagonal Toeplitz matrix with the diagonal element equal to 2 and the off-diagonal element equal to 1, (b) matrices obtained from the matrix defined in (a) by varying the diagonal elements, and (c) random tridiagonal matrices. The order of the test matrices is $N = pn = qm$, where $p = 32$ is the number of the "physical" processors actually available and q is the number of "logical" processors (see section 5). Table 2 gives the running times for each of the following stages of the algorithm.

1. D—Logical Diagonalization.
2. I—Bottom-Up and Top-Down stages performed by the logical processors within any physical processor.
3. E—Bottom-Up and Top-Down stages performed by the physical processors.
4. C—Logical Factorization.
5. S—Sequential algorithm.

TABLE 2
Computational examples on the CM5, $q = 2^{16}$.

N	$3 * 2^{24}$	2^{25}	2^{24}
m	$3 * 2^8$	2^9	2^8
D	1.444	0.963	0.480
I	0.033	0.033	0.033
E	0.014	0.014	0.014
C	0.562	0.372	0.186
total	2.054	1.376	0.713
S	2704.12	1791.73	880.85
speedup	1316	1296	1235

Clearly, the speedup is larger than $p/4$. Besides the additional parallelization due to having four vector units, we gain a factor of ~ 40 due to vectorization.

In Table 3 we depict similar results for matrices of smaller size. The decrease in performance is due to shorter vector length and the increased effect of communication overheads. Thus, as the matrix size becomes smaller we should consider as a better strategy using fewer processors. (We are not able to report on such experiments due to the fixed system partition in ICSI.) The performance we observed suggests that we should use vector sizes ≥ 128 and blocks of order ≥ 64, so that for $N = 4 * 2^{13}p$ we should use p processors.

8. Error analysis. The main result of the a priori analysis (see [7]) is that the pivots $\hat{x}_{nk}$ and $\hat{x}'_{nk}$, computed by processor k at the end of stages (ii) and (iii), respectively, satisfy

$$|\hat{x}_{nk}/\hat{x}'_{nk}| = (1 + \eta), \qquad |\eta| = O((n + \log p)\hat{\theta}),$$

where $\hat{\theta}$ represents the input precision, provided we use some higher precision $\theta < \hat{\theta}$ in the computation. To appreciate the significance of this result we proceed with the following.

TABLE 3
Smaller size matrices.

N	2^{20}	2^{19}	2^{18}
q	2^{14}	2^{13}	2^{13}
m	2^6	2^6	2^5
D	0.037	0.022	0.011
I	0.012	0.007	0.007
E	0.014	0.014	0.014
C	0.016	0.010	0.005
total	0.079	0.053	0.037
S	50.43	24.78	12.36
speedup	638	468	334

A posteriori error analysis. Consider stage (iii) of the parallel algorithm, i.e., the actual transformation applied to the matrix. As in (7), let B denote the block assigned to a given processor p, and let x_0 be the pivot computed (during the first two stages) by processor $p-1$. Processor p computes the following recurrences (see section 4):

$$z_i = c_i/x_{i-1}, \quad x_i = a_i - z_i, \quad i = 1, \ldots, n.$$

Taking the rounding errors into account, we have

$$\hat{z}_i = \hat{c}_i/\hat{x}_{i-1}, \quad \hat{c}_i = c_i(1+\epsilon_i), \quad |\epsilon_i| \le \theta,$$
$$\hat{x}_i = \hat{a}_i - \hat{z}_i, \quad \hat{a}_i = a_i(1+\eta_i), \quad |\eta_i| \le \theta,$$

where $\hat{x}_0$ is the computed pivot. In this analysis there is a discrepancy; i.e., $\hat{x}_n$ is not the same as the pivot transmitted to processor $p+1$. To fix this problem, define $\hat{x}'_0$ as the pivot computed by processor $p-1$ by the end of stage (iii); then the first step above can be written as $\hat{z}_1 = (c_1/\hat{x}'_0)(1+\epsilon_1)(\hat{x}'_0/\hat{x}_0) = \hat{c}_1/\hat{x}'_0$, where $\hat{c}_1 = (1+\epsilon'_1)c_1$, and $(1+\epsilon'_1) = (1+\epsilon_1)(\hat{x}'_0/\hat{x}_0)$.

The above argument shows that the solution computed by our parallel algorithm is the exact solution of a system in which the first off-diagonal elements of each block are further perturbed (with respect to the classical sequential algorithm) by the factor $\hat{x}'_0/\hat{x}_0$. Hence, when $|(\hat{x}'_0 - \hat{x}_0)/\hat{x}_0| = O(\theta)$ the algorithm is componentwise stable in the backward sense, and this is confirmed by the a priori analysis.

We found experimentally that $|(\hat{x}'_0 - \hat{x}_0)/\hat{x}_0|$ is relatively small even using standard double precision on very ill-conditioned matrices; see Table 4. The table contains results related to three different kinds of tests: (i) Test 1: random diagonally dominant matrices; (ii) Test 2: the Toeplitz tridiagonal symmetric matrix with $a_{ii} = 2$ and $a_{i+1,i} = a_{i,i+1} = 1$; (iii) Test 3: random tridiagonal matrices. We have added an appropriate shift to the diagonal elements to assure positive definiteness of the perturbed matrix.

9. Further work. The efficiency of the LR scheme, accelerated with our algorithm in the decomposition stage, should be compared with other algorithms for the computation of the eigensystem of tridiagonal symmetric matrices, notably QR [6] and divide-and-conquer algorithms [4, 10, 13].

Possible generalizations of this work include the cases of block tridiagonal and band matrices. In fact, for both kinds of matrices the algorithmic framework appears

TABLE 4
Error analysis for $N = 2^{23}$ and $m = 2^8$. Each test column gives the number of correct digits in the computed factorization as produced by the a posteriori error bound.

Shift	Test 1	Test 2	Test 3
10^{-4}	14	15	16
10^{-8}	12	14	16
10^{-12}	12	14	8
10^{-14}	12	14	6

to be essentially the same. In addition, it is possible to apply similar ideas to the development of a parallel band version of the QR algorithm.

Acknowledgments. The authors are indebted to the referees for a number of suggestions that helped to improve the quality and the clarity of the paper. The first author is also indebted to Prof. F.T. Luk for his encouragement and advice.

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, 1992.

[2] I. BAR-ON, *A practical parallel algorithm for solving band symmetric positive definite systems of linear equations*, ACM Trans. Math. Software, 13 (1987), pp. 323–332.

[3] I. BAR-ON, *Fast Parallel LR and QR Algorithms for Symmetric Band Matrices*, Tech. report 749, Technion, Computer Science Department, Haifa, Israel, 1992.

[4] I. BAR-ON, *A New Divide and Conquer Parallel Algorithm for Computing the Eigenvalues of a Symmetric Tridiagonal Matrix*, Tech. report 832, Technion, Computer Science Department, Haifa, Israel, 1994.

[5] I. BAR-ON, *Interlacing properties for tridiagonal symmetric matrices with applications to parallel computing*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 548–562.

[6] I. BAR-ON AND B. CODENOTTI, *A fast and stable parallel QR algorithm for symmetric tridiagonal matrices*, Linear Algebra Appl., 220 (1995), pp. 63–96.

[7] I. BAR-ON, B. CODENOTTI, AND M. LEONCINI, *A Fast Parallel Cholesky Decomposition Algorithm for Tridiagonal Symmetric Matrices*, Tech. report TR-95-006, International Computer Science Institute, Berkeley, CA, 1995.

[8] I. BAR-ON AND M. LEONCINI, *Fast and reliable parallel solution of bidiagonal systems*, SIAM J. Numer. Anal., 1995, submitted.

[9] J. BARLOW AND J. DEMMEL, *Computing accurate eigensystems of scaled diagonally dominant matrices*, SIAM J. Numer. Anal., 27 (1990), pp. 762–791.

[10] J. CUPPEN, *A divide and conquer method for the symmetric eigenproblem*, Numer. Math., 36 (1981), pp. 177–195.

[11] J. DEMMEL AND W. KAHAN, *Accurate singular values of bidiagonal matrices*, SIAM J. Sci. Stat. Comput., 11 (1990), pp. 873–912.

[12] J. J. DONGARRA AND A. H. SAMEH, *On some parallel banded system solvers*, Parallel Comput., 1 (1984), pp. 223–235.

[13] J. J. DONGARRA AND D. SORENSEN, *A fully parallel algorithm for the symmetric eigenvalue problem*, SIAM J. Sci. Stat. Comput., 8 (1987), pp. s139–s154.

[14] Z. DRMAČ, M. OMLADIČ, AND K. VESELIĆ, *On the perturbation of the Cholesky factorization*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1319–1332.

[15] K. V. FERNANDO AND B. PARLETT, *Accurate singular values and differential qd algorithms*, Numer. Math., 67 (1994), pp. 191–229.

[16] J. FRANCIS, *The QR transformation, part* I, Comput. J., 4 (1961), pp. 265–271.

[17] J. FRANCIS, *The QR transformation, part* II, Comput. J., 4 (1962), pp. 332–345.

[18] J. GRAD AND E. ZAKRAJŠEK, *LR algorithm with Laguerre shift for symmetric tridiagonal matrices*, Comput. J., 15 (1972), pp. 268–270.

[19] S. L. JOHNSSON, *Solving tridiagonal systems on ensemble architectures*, SIAM J. Sci. Stat. Comput., 8 (1987), pp. 354–392.

[20] B. N. Parlett, *Laguerre's method applied to the matrix eigenvalue problem*, Math. Comput., 18 (1964), pp. 464–485.
[21] H. Rutishauser, *Solution of eigenvalue problems with the LR transformation*, Nat. Bur. Standards, AMS, 49 (1958), pp. 47–81.
[22] H. Rutishauser and H. Schwarz, *The LR transformation method for symmetric matrices*, Numer. Math., 5 (1963), pp. 273–289.
[23] A. Sameh and D. Kuck, *On stable parallel linear system solver*, J. Assoc. Comput. Mach., 25 (1978), pp. 81–91.
[24] H. S. Stone, *An efficient parallel algorithm for the solution of a tridiagonal system of equations*, J. Assoc. Comput. Mach., 20 (1975), pp. 27–38.
[25] H. S. Stone, *Parallel tridiagonal equation solver*, ACM Trans. Math. Software, 1 (1975), pp. 289–307.
[26] P. Swarztrauber, *A parallel algorithm for solving general tridiagonal equations*, Math. Comput., 33 (1979), pp. 185–199.
[27] Thinking Machine Corporation, *The Connection Machine CM-5 Technical Summary*, 1991.
[28] H. Wang, *A parallel method for tridiagonal equations*, ACM Trans. Math. Software, 7 (1981), pp. 170–183.
[29] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, UK, 1965. Reprinted in Oxford Science Publications, 1988.
[30] J. H. Wilkinson, *A priori error analysis of algebraic processes*, in Proc. International Congress Math., Moscow: Izdat, 1968, pp. 629–639.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, pp. 419–434, April 1997

012

A STABILIZED QMR VERSION OF BLOCK BICG*

V. SIMONCINI†

Abstract. Block BICG (BBICG) is an appealing method for solving $AX = B$ with $A \in \mathbb{R}^{n\times n}$ and $X, B \in \mathbb{R}^{n\times s}$. Because of its short-term recurrence form, memory allocation and computational cost do not depend on additional parameters. Unfortunately, loss of orthogonality prevents convergence in many cases.

We present a new version of the algorithm that generates blocks of vectors that are vectorwise A-biorthogonal; moreover, a near-breakdown safeguard strategy inside the block stabilizes the computation of the coefficients. In order to smooth the possibly erratic behavior of the residual norm curve, the approximate solution is determined using a block QMR procedure. The new method considerably improves the robustness of the original algorithm, showing very good performance on dense or preconditioned matrices over both BBICG and the single right-hand side solver coupled two-term QMR method applied on each system.

Key words. Krylov subspace, block iterative methods, two-sided Gram–Schmidt, multiple right-hand sides, large linear systems

AMS subject classification. 65F10

PII. S0895479894264673

1. Introduction.

Some application problems require the solution of systems of linear equations with the same coefficient matrix but different right-hand sides; see the references in [30]. When the dimension n of the problem is small, a convenient way to solve the given systems consists of decomposing the coefficients matrix into two simpler factors and then solving the corresponding subproblem for all available right-hand sides [14, 7]. Direct methods can also be advantageous if the right-hand sides are not all simultaneously available. However, for n large this approach can be prohibitively expensive both in terms of memory and computational cost, and so iterative schemes become appealing. We are thus looking at the problem

$$AX = B, \qquad A \in \mathbb{R}^{n\times n}, X, B \in \mathbb{R}^{n\times s} \tag{1.1}$$

with A nonsymmetric, n large, and $s \ll n$. Given a first guess $X_0 \in \mathbb{R}^{n\times s}$, we would like to determine an approximation $X_m = X_0 + \mathcal{Z}_m$ of X with $\mathcal{Z}_m$ belonging to the block Krylov subspace $\mathbb{K}_m(A, R_0) = \text{span}\{R_0, AR_0, \ldots, A^{m-1}R_0\}$, where $R_0 = B - AX_0$. In theory, the presence of blocks allows the computation of a good approximation X_m with m smaller than if a single Krylov subspace were to be built for each system; in practice, however, this is not always the case [31].

Natural implementations of methods that approximate X using a block Krylov subspace are generalizations of single right-hand side solvers; available algorithms either explicitly compute a basis of $\mathbb{K}_m(A, R_0)$ (such as BGMRES [32]) or use a short-term recurrence for generating implicitly linearly independent elements of $\mathbb{K}_m(A, R_0)$ (such as BBICG [23]). The choice between the two classes of methods is based on several parameters, among which is the availability of the transpose of A. BGMRES

* Received by the editors March 16, 1994; accepted for publication (in revised form) by R. Freund May 31, 1996. Part of this work was done while the author was visiting CSRD, University of Illinois at Urbana-Champaign in November–December 1993. This research was supported by DARPA under a subcontract from the University of Minnesota grant DARPA/NIST 60NANB2D1272.

http://www.siam.org/journals/simax/18-2/26467.html

† Dipartimento di Fisica, Università di Bologna, Bologna, Italy. Current address: Istituto di Analisi Numerica, CNR, Via Abbiategrasso 209, 27100 Pavia, Italy (val@dragon.ian.pv.cnr.it).

requires the computation and storage of an orthogonal basis. Therefore, memory limitations and high computational cost force restarting; the method can then experience stagnation or slow convergence [30]. Due to a biorthogonality condition, BBICG exploits a short-term recurrence and does not require the explicit generation of an orthogonal basis. As an important consequence, BBICG in general avoids restart. On the other hand, round-off errors quickly destroy the linear independence of the generated blocks, affecting the convergence of the method [30].

Due to the high computational cost per iteration, block methods are advantageous when (i) the total number of iterations decreases considerably with respect to the single right-hand side solver or (ii) a moderately lower number of iterations is accompanied by a high cost of matrix-vector operations with the coefficients matrix A. Consequently, the block approach can be effective when dealing with dense or preconditioned matrices [18, 29, 20].

It is well known that robustness is related to a stable computation of the iteration coefficients. In this paper we propose a new version of BBICG, reviewed in section 2, in which the computation of the recurrence coefficients is modified. Four sets of matrices are introduced,

$$\mathcal{Q}_k = [Q_0, \ldots, Q_k], \quad \tilde{\mathcal{Q}}_k = [\tilde{Q}_0, \ldots, \tilde{Q}_k],$$
$$\mathcal{S}_k = [S_0, \ldots, S_k], \quad \tilde{\mathcal{S}}_k = [\tilde{S}_0, \ldots, \tilde{S}_k],$$

that are mutually orthogonal; that is, they satisfy $\tilde{\mathcal{Q}}_k^T \mathcal{Q}_k = \mathcal{D}_\mathcal{Q}$ and $\tilde{\mathcal{S}}_k^T A \mathcal{S}_k = \mathcal{D}_\mathcal{S}$ with $\mathcal{D}_\mathcal{Q}$ and $\mathcal{D}_\mathcal{S}$ diagonal matrices. Note that in the original implementation of BBICG both matrices $\mathcal{D}_\mathcal{Q}$ and $\mathcal{D}_\mathcal{S}$ are only required to be block diagonal. We provide experimental evidence that full biorthogonality, rather than only block biorthogonality, improves the robustness of the original method. We would like to mention that the idea of computing vectorwise orthogonal vectors was used by Ruhe in the symmetric eigenvalue problem for stabilizing the Lanczos recurrence [26].

During the recurrence process, loss of rank may occur on matrices Q_k, S_k or $\tilde{Q}_k, \tilde{S}_k$. If this is the case, the exact solution of a related problem has been determined. In order to deflate the block iterates, we restart the algorithm with the new full rank bases. Another strategy would consist of continuing the recurrence process with the deflated blocks, while keeping track in some way of the deflated vectors to maintain biorthogonality. Due to the difficulties that the implementation of such a strategy would encounter on the block algorithm, we have decided to discard this approach.

In section 3.1 and section 3.2 the tools used in the implementation of the new algorithm, namely, the Gram–Schmidt biorthogonalization and the block quasi-minimal residual procedure, will be described. In section 4 the new recurrences and the resulting algorithm are derived. The approximate solution is then determined via a block smoothing residual technique based on the single right-hand side procedure QMR [11]. Some stability issues are discussed in section 5. In particular, it is shown that in some cases almost-singularity of $\mathcal{D}_\mathcal{S}$ or $\mathcal{D}_\mathcal{Q}$ can be overcome by simply permuting the columns of the corresponding iteration matrices. In section 6 we show that for a certain class of matrices the new method (QMR-MBCG) performs considerably better than both the original BBICG and the corresponding single right-hand side solver. Some applications are given in section 6.2.

The following notation will be used. The Frobenius norm for matrices and vectors will be denoted by $\|\cdot\| \equiv \|\cdot\|_F$. Matrix $0_{n,s}$ is the zero matrix in $\mathbb{R}^{n\times s}$, abbreviated with 0_s for $n = s$; matrix $I_{n,s}$ is the principal submatrix of the identity matrix in $\mathbb{R}^n$; and I_n will denote the corresponding square matrix. Moreover,

$E_i := [0_s, \ldots, I_s, \ldots, 0_s]^T$ with I_s at the ith block position; the total size of E_i will be made clear from the context. We will denote by $\mathrm{diag}(d_1, \ldots, d_m)$ a block diagonal matrix of size m and diagonal elements $d_i \in \mathbb{R}^{s\times s}$. Vector $X_{:,i}$ is the ith column of matrix X, for which $|X|$ is the elementwise absolute value.

2. Review of block BICG. The block BICG algorithm, originally proposed by O'Leary [23] as a generalization of the biconjugate gradient method [19], computes two sets of *direction* matrices $\{P_0, \ldots, P_{m-1}\}$ and $\{\tilde{P}_0, \ldots, \tilde{P}_{m-1}\}$ that span the subspaces $\mathbb{K}_m(A, R_0)$ and $\mathbb{K}_m(A^T, \tilde{R}_0)$, respectively, where $\tilde{R}_0$ is a chosen additional matrix. Letting $P_0 = R_0$ and $\tilde{P}_0 = \tilde{R}_0$, the iterates are computed using the following recursions:

$$P_{i+1} = R_{i+1} + P_i\beta_i, \qquad \tilde{P}_{i+1} = \tilde{R}_{i+1} + \tilde{P}_i\tilde{\beta}_i, \tag{2.1}$$

$$R_{i+1} = R_i - AP_i\alpha_i, \qquad \tilde{R}_{i+1} = \tilde{R}_i - A^T\tilde{P}_i\tilde{\alpha}_i, \tag{2.2}$$

where the coefficients are defined as

$$\beta_i = (\tilde{R}_i^T R_i)^{-1}\tilde{R}_{i+1}^T R_{i+1}, \quad \tilde{\beta}_i = (R_i^T\tilde{R}_i)^{-1}R_{i+1}^T\tilde{R}_{i+1}, \tag{2.3}$$

$$\alpha_i = (\tilde{P}_i^T AP_i)^{-1}\tilde{R}_i^T R_i, \quad \tilde{\alpha}_i = (P_i^T A^T\tilde{P}_i)^{-1}R_i^T\tilde{R}_i. \tag{2.4}$$

In practice, the matrices $P_i, \tilde{P}_i$ can be computed so as to have orthogonal columns [23, section 2]. Moreover, the iterates satisfy

$$\tilde{R}_j^T R_i = 0, \quad i > j, \qquad \text{biorthogonality condition,} \tag{2.5}$$

$$\tilde{P}_j^T AP_i = 0, \quad i > j, \qquad A\text{-biorthogonality condition.} \tag{2.6}$$

In finite arithmetic computation, these relations may not be satisfied even for $|i - j|$ small (cf. section 6.1). Indeed, the computation of the coefficients in (2.3) and (2.4) requires the solution of $s\times s$ systems with possibly ill-conditioned coefficients matrices. This will affect the computation of the next iterates.

The algorithm is also characterized by the erratic behavior of the residual norm. This problem can be overcome using a quasi-minimization procedure; see [9, 11] and the references therein for a general treatment of the QMR approach in Lanczos methods. The derivation of a block QMR scheme is described in section 3.2.

Finally, the residuals R_i and $\tilde{R}_i$ are involved in the computation. This may cause some computational problems, especially when the associated system is close to convergence; see section 6.2.

3. Some tools.

3.1. The two-sided modified Gram–Schmidt procedure. It was observed in [24] that *the Gram–Schmidt procedure does not require an inner product.* Indeed, given $x, y \in \mathbb{R}^n$, the bilinear form $(x, y) := x^T y$ can be used for generating sets of biorthogonal vectors $[x_1, \ldots, x_s]$ and $[y_1, \ldots, y_s]$, thus satisfying $(x_i, y_j) = 0$ and $(x_i, y_i) \neq 0$ for $i, j = 1, \ldots, s$, $j \neq i$; see also [5, 1]. Obviously, it may be that for certain x_i, y_i $(x_i, y_i) = 0$; in this case the biorthogonalization will not be successful. The basic procedure looks as follows.

FUNCTION $[Q, \xi, \tilde{Q}, \tilde{\xi}, \Omega] = \mathrm{MGS}(X, Y)$.

(a) for $i = 1, s$

(b) $\quad Q_{:,i} = X_{:,i}$

(c) $\quad \tilde{Q}_{:,i} = Y_{:,i}$

(d) for $j = 1, i-1$
(e) $\xi_{j,i} = \Omega_{j,j}^{-1} \tilde{Q}_{:,j}^T Q_{:,i}, \qquad Q_{:,i} = Q_{:,i} - Q_{:,j}\xi_{j,i}$
(f) $\tilde{\xi}_{j,i} = \Omega_{j,j}^{-T} Q_{:,j}^T \tilde{Q}_{:,i}, \qquad \tilde{Q}_{:,i} = \tilde{Q}_{:,i} - \tilde{Q}_{:,j}\tilde{\xi}_{j,i}$
(g) end
(h) $\xi_{i,i} = \|Q_{:,i}\|, \qquad Q_{:,i} = Q_{:,i}\xi_{i,i}^{-1}$
(i) $\tilde{\xi}_{i,i} = \|\tilde{Q}_{:,i}\|, \qquad \tilde{Q}_{:,i} = \tilde{Q}_{:,i}\tilde{\xi}_{i,i}^{-1}$
(l) $\Omega_{i,i} = \tilde{Q}_{:,i}^T Q_{:,i}$
(m) end

Given $X, Y \in \mathbb{R}^{n\times s}$, function MGS determines matrices $Q, \tilde{Q}$ of the same size as X, Y and normal columns such that $\tilde{Q}^T Q = \Omega$ diagonal and $Y = \tilde{Q}\tilde{\xi}$, $X = Q\xi$ with $\xi, \tilde{\xi}$ $s \times s$ upper triangular matrices; see [24]. The computational cost of MGS is about $4ns^2 + 8ns$ floating point operations (flops).

Given $A \in \mathbb{R}^{n\times n}$ and matrices $Y, Z \in \mathbb{R}^{n\times s}$, a similar problem consists of finding $S, \tilde{S}$ such that $\tilde{S}^T A S = \Omega$ and $Z = S\xi, Y = \tilde{S}\tilde{\xi}$ with Ω, ξ and $\tilde{\xi}$ as before. MGS can be adapted to handle this case; we will call this version MGS.A. The input matrices of MGS.A are AZ, Y, Z, where AZ plays the role of X in MGS. Function MGS.A is obtained from MGS by adding the following lines:

(c′) $T_{:,i} = Z_{:,i}$
(f′) $T_{:,i} = T_{:,i} - T_{:,j}\xi_{j,i}$
(i′) $T_{:,i} = T_{:,i}\xi_{i,i}^{-1}$

thus transforming (AZ, Y, Z) into the triple $(Q, \tilde{Q}, T)$ such that $\tilde{Q}^T A T = \Omega$ diagonal, $Q = AT$, and $(\tilde{Q}\tilde{\xi})^T Q\xi = Y^T AZ$ for a total cost of $5ns^2 + 9ns$ flops. As a by-product, the routine implicitly determines $Q = AT$, which does not need to be computed again.

As already mentioned, due to the bilinear form $(x, y) = x^T y$, both algorithms can fail to generate Q and $\tilde{Q}$ if, for any j, $\Omega_{j,j} = 0$; the treatment of this important problem is postponed until section 5.

3.2. Smoothing the curve of the residual norm. The QMR procedure was first introduced by Freund and Nachtigal for smoothing the residual convergence curve of the Lanczos method for solving non-Hermitian linear systems [11]. We will be referring to QMR not as a system solver, but as a smoothing technique applicable to the residual recurrence of a Krylov subspace solver characterized by a highly oscillating convergence curve; an analysis of the QMR smoothing can be found in [6, 33]. The procedure is described below for the recurrence iterates generated by BBICG. After i iterations, BBICG has generated blocks $\mathcal{Q}_i, \mathcal{S}_i \in \mathbb{R}^{n\times(i+1)s}$ satisfying

$$A\mathcal{S}_{i-1}\Gamma_{i-1} = \mathcal{Q}_i\mathcal{T}_i \tag{3.1}$$

with $\Gamma_i = \mathrm{diag}(\alpha_0, \ldots, \alpha_i)$ and $\mathcal{T}_i \in \mathbb{R}^{(i+1)s\times is}$ block bidiagonal matrix, $\mathcal{T}_i = (T_{k,j})$. The approximate solution is written as $X_i = X_0 + \mathcal{S}_{i-1}\Gamma_{i-1}y$, $y \in \mathbb{R}^{is\times s}$, and if $R_0 = Q_0\xi_0$, the residual $R_i = B - AX_i$ becomes $R_i = \mathcal{Q}_i(E_1\xi_0 - \mathcal{T}_i y)$. The block quasi-minimization procedure consists of minimizing the quantity

$$\min_{y\in\mathbb{R}^{is\times s}} \|\mathcal{W}_i(E_1\xi_0 - \mathcal{T}_i y)\|, \tag{3.2}$$

where $\mathcal{W}_i$ is a weight matrix that allows us to compute a simple approximation of the residual norm [29]. If, for instance, the columns of $\mathcal{Q}_i$ have unit norm, the computation of $\mathcal{W}_i$ can be avoided. Problem (3.2) is solved by doing a QR decomposition of $\mathcal{T}_i$ with

factors Y_i and U_i such that $Y_i^T Y_i = I$ and U_i is upper triangular. Due to the banded structure of $\mathcal{T}_i$, this can be carried out via Givens rotations in an incremental manner. Moreover, matrix $\mathcal{S}_{i-1}U_i^{-1}$ can be computed with a short-term recurrence formula so that $X_{i+1} = X_i + Z_{i+1}\eta_i$, where $\mathcal{S}_{i-1}U_i^{-1} = [Z_0, \ldots, Z_i]$. Note that the elements η_i of $\eta := Y^T E_1 \xi_0$ are also updated iteratively. The final scheme is as follows.

FUNCTION $[Z_{i+1}, \eta_i, \tilde{\eta}_{i+1}, G_i] = \text{MINIM}(T_{i+1,i}, \gamma_i, \tilde{\eta}_i, Z_i, S_i, G_{i-1})$.

- Apply the rotations of the previous step (matrix G_{i-1})
- Generate new rotations matrix G_i
- Update η with the new rotations
- Generate Z_{i+1} with the new columns of $\mathcal{T}_i$

We refer to [29] for a detailed description of the procedure in block form. The major computational cost of the process is $6ns^2 + ns + 2s^3$ flops.

It is evident that the procedure only acts on the approximate solution iterate by doing an appropriate decomposition of $\mathcal{T}_i$, while the columns of $\mathcal{Q}_i$ and $\mathcal{S}_i$ are generated by the underlying method. Therefore, the robustness properties of the method remain unchanged.

4. A new recurrence. We next describe how to generate the iterates of the sets $\mathcal{Q}_k, \tilde{\mathcal{Q}}_k, \mathcal{S}_k$, and $\tilde{\mathcal{S}}_k$ that are vectorwise biorthogonal. We will use the procedure introduced in section 3.1 to stabilize the computation of the iteration parameters. This will give the recurrence formulas for the new algorithm.

Assume that for $k = 0, \ldots, i$ there exist iterates $Q_k, \tilde{Q}_k, L_k, \tilde{L}_k$ satisfying $R_0 = Q_0\xi_0$, $\tilde{R}_0 = \tilde{Q}_0\tilde{\xi}_0$, $\tilde{Q}_k^T Q_k = \Omega_k$ with Ω_k diagonal and $R_0 = L_0\tau_0$, $\tilde{R}_0 = \tilde{L}_0\tilde{\tau}_0$ such that $\Pi_0 = \tilde{L}_0^T A L_0$ is diagonal. Let $S_k, \tilde{S}_k$ be the A-biorthogonal form of $L_k, \tilde{L}_k$, respectively; that is,

$$L_k = S_k\tau_k, \qquad \tilde{L}_k = \tilde{S}_k\tilde{\tau}_k \qquad \text{with} \qquad \tilde{S}_k^T A S_k = \Pi_k \quad \text{diagonal.} \tag{4.1}$$

We also assume that iterates $Q_k, \tilde{Q}_k$ and $S_k, \tilde{S}_k$ satisfy (2.5) and (2.6). We want to determine short-term recurrence relations to compute the next iterates $Q_{i+1}, \tilde{Q}_{i+1}$, S_{i+1}, and $\tilde{S}_{i+1}$.

The first step consists of replacing the recursions (2.2) with iterations in Q_i, $\tilde{Q}_i$ and L_i, $\tilde{L}_i$. We seek matrices $\sigma_i, \tilde{\sigma}_i$ that force mutual orthogonality of the new blocks

$$g_{i+1} = Q_i - AL_i\sigma_i, \qquad \tilde{g}_{i+1} = \tilde{Q}_i - A^T\tilde{L}_i\tilde{\sigma}_i \tag{4.2}$$

with respect to the previous iterates. Hence we define

$$\sigma_i = (\tilde{L}_i^T A L_i)^{-1}\tilde{Q}_i^T Q_i, \qquad \tilde{\sigma}_i = (L_i^T A^T \tilde{L}_i)^{-1} Q_i^T \tilde{Q}_i. \tag{4.3}$$

From (4.1) it follows that $\tilde{L}_i^T A L_i = \tilde{\tau}_i^T \Pi_i \tau_i$. Moreover, the definition of Q_i and $\tilde{Q}_i$ gives[1]

$$\sigma_i = \tau_i^{-1}\Pi_i^{-1}\tilde{\tau}_i^{-T}\Omega_i, \qquad \tilde{\sigma}_i = \tilde{\tau}_i^{-1}\Pi_i^{-T}\tau_i^{-T}\Omega_i^T.$$

Using (4.1) once more, the recursions (4.2) can be written as

$$g_{i+1} = Q_i - AS_i\Pi_i^{-1}\tilde{\tau}_i^{-T}\Omega_i, \qquad \tilde{g}_{i+1} = \tilde{Q}_i - A^T\tilde{S}_i\Pi_i^{-T}\tau_i^{-T}\Omega_i^T. \tag{4.4}$$

[1] The notation Ω^T, Π^T for Ω, Π diagonal is used for keeping track of transposition when using the safeguard procedure of section 5.1.

We determine the new iterates $Q_{i+1}, \tilde{Q}_{i+1}$ so that biorthogonality inside the block is preserved, that is, using the two-sided Gram–Schmidt decomposition,

$$g_{i+1} = Q_{i+1}\xi_{i+1}, \qquad \tilde{g}_{i+1} = \tilde{Q}_{i+1}\tilde{\xi}_{i+1} \quad \text{with} \quad \tilde{Q}_{i+1}^T Q_{i+1} = \Omega_{i+1}. \tag{4.5}$$

Relations (4.4) and (4.5) replace the original ones in (2.2). However, iterates $R_{i+1}, \tilde{R}_{i+1}$ can be recovered by observing that

$$R_i = Q_i\xi_i\xi_{i-1}\dots\xi_0, \qquad \tilde{R}_i = \tilde{Q}_i\tilde{\xi}_i\tilde{\xi}_{i-1}\dots\tilde{\xi}_0. \tag{4.6}$$

Recursions involving the direction matrices P_i and $\tilde{P}_i$ change accordingly. The substitution of (4.6) and (4.5) in (2.1) gives

$$P_{i+1} = Q_{i+1}\xi_{i+1}\cdots\xi_0 + P_i(\xi_i\cdots\xi_0)^{-1}\Omega_i^{-1}\tilde{\xi}_{i+1}^T\Omega_{i+1}\xi_{i+1}\cdots\xi_0.$$

An analogous relation holds for $\tilde{P}_{i+1}$. Let $L_i := P_i(\xi_i\cdots\xi_0)^{-1}$, $\tilde{L}_i := \tilde{P}_i(\tilde{\xi}_i\cdots\tilde{\xi}_0)^{-1}$. Thus, the recurrences

$$L_{i+1} = Q_{i+1} + L_i\Omega_i^{-1}\tilde{\xi}_{i+1}^T\Omega_{i+1}, \qquad \tilde{L}_{i+1} = \tilde{Q}_{i+1} + \tilde{L}_i\Omega_i^{-T}\xi_{i+1}^T\Omega_{i+1}^T \tag{4.7}$$

complete the description of the new scheme. The iterates $S_{i+1}, \tilde{S}_{i+1}$ are generated by forcing A-biorthogonality between the new matrices $L_{i+1}, \tilde{L}_{i+1}$ in (4.7).

Using (4.6) and the definition of $L_i, \tilde{L}_i$, the original BBICG recurrence $X_{i+1} = X_i + P_i\alpha_i$ becomes

$$X_{i+1} = X_i + S_i\Pi_i^{-1}\tilde{\tau}_i^{-T}\Omega_i\chi_i \quad \text{with} \quad \chi_i = \xi_i\chi_{i-1}, \quad \chi_0 = \xi_0.$$

An approximate solution with a smoother residual norm can be obtained using the quasi-minimization technique of section 3.2. The ith block column in the matricial equality (3.1) thus becomes

$$AS_i\gamma_i = [\dots, Q_{i-1}, Q_i,\, g_{i+1}] \begin{bmatrix} \vdots \\ 0_s \\ I_s \\ -\xi_{i+1} \end{bmatrix}.$$

The principal steps of the new algorithm QMR-MBCG are summarized below, where the acronym stands for the quasi-minimum residual norm version of modified BBICG. For the system (1.1) right preconditioning has been used, which allows the minimization of the unpreconditioned residual.

ALGORITHM X=QMR-MBCG$(A, M, B, X_0, \tilde{R}, \varepsilon)$.
$X = X_0, R = B - AX_0$
$Z = 0_{n,s}$
$[Q_0, \xi_0, \tilde{Q}_0, \tilde{\xi}_0, \Omega_0] = \text{MGS}(R, M^{-T}\tilde{R})$
$L_0 = R, \quad \tilde{L}_0 = \tilde{R}$
$[S_0, \tau_0, AM^{-1}S_0, \tilde{S}_0, \tilde{\tau}_0, \Pi_0] = \text{MGS.A}(AM^{-1}L_0, \tilde{L}_0, L_0)$
$G_{-1} = I_{2s}, \quad \tilde{\eta}_0 = \xi_0$
for $i = 0, 1, \dots,$
$\gamma_i = \Pi_i^{-1}\tilde{\tau}_i^{-T}\Omega_i, \qquad \tilde{\gamma}_i = \Pi_i^{-T}\tau_i^{-T}\Omega_i^T$
$g_{i+1} = Q_i - AS_i\gamma_i, \quad \tilde{g}_{i+1} = \tilde{Q}_i - M^{-T}A^T\tilde{S}_i\tilde{\gamma}_i$
$[Q_{i+1}, \xi_{i+1}, \tilde{Q}_{i+1}, \tilde{\xi}_{i+1}, \Omega_{i+1}] = \text{MGS}(g_{i+1}, \tilde{g}_{i+1})$

$[Z, \eta_i, \tilde{\eta}_{i+1}, G_i] = \text{MINIM}(\xi_{i+1}, \gamma_i, \tilde{\eta}_i, Z, S_i, G_{i-1})$
$X = X + Z\eta_i$
If X good enough then stop
$\delta_i = \Omega_i^{-1}\tilde{\xi}_{i+1}^T\Omega_{i+1}, \qquad \tilde{\delta}_i = \Omega_i^{-T}\xi_{i+1}^T\Omega_{i+1}^T$
$L_{i+1} = Q_{i+1} + L_i\delta_i, \qquad \tilde{L}_{i+1} = \tilde{Q}_{i+1} + \tilde{L}_i\tilde{\delta}_i,$
$[S_{i+1}, \tau_{i+1}, AM^{-1}S_{i+1}, \tilde{S}_{i+1}, \tilde{\tau}_{i+1}, \Pi_{i+1}] = \text{MGS.A}(AM^{-1}L_{i+1}, \tilde{L}_{i+1}, L_{i+1})$
end

The biorthogonalization functions MGS and MGS.A are as described in section 3.1. We note that MGS generates biorthonormal matrices $Q, \tilde{Q}$; thus the weight matrix $\mathcal{W}_k$ in (3.2) can be taken to be the identity matrix. Convergence can be checked at each step using the upper bound [29]

$$\|B - AX_i\| \leq s\sqrt{i+1}\|\tilde{\eta}_{i+1}\|.$$

The residual of the underlying BBICG can be recovered by updating formula (4.6), while the true residual could be determined by adding a recurrence relation [11, 29]. A theoretical analysis of convergence properties of block Krylov subspace methods was recently proposed in [31], where it was shown that standard results for $s = 1$ can be naturally generalized to $s \geq 1$ by using matrix polynomials; see also [29] for results concerning the block QMR approach.

We also remark that when A is symmetric the iteration recurrence is analogous to that of the modified block CG algorithm introduced in [1]. Moreover, QMR-MBCG can be seen as a generalization to multiple systems of the coupled two-term QMR method proposed in [12].

5. On the robustness of modified BBICG. It is known that loss of orthogonality affects the performance of BICG and, in general, of Lanczos methods [10, 25]. The block form of BBICG further exacerbates this problem, precluding the convergence of the method in many cases [30]. The new implementation, QMR-MBCG, is characterized by local biorthogonality with high accuracy, while maintaining global biorthogonality with low accuracy. Incidentally, global properties seem to be better preserved in the new scheme. The two Gram–Schmidt procedures also yield a more stable computation of the recurrence parameters. Indeed, inversions are carried out with diagonal or triangular matrices, which are usually very accurate (cf. section 6.1).

Analogously to the single right-hand side case, BBICG may break down if, at any iteration i, $\tilde{R}_i^T R_i$ or $\tilde{P}_i^T A P_i$ is singular with $R_i, \tilde{R}_i, \tilde{P}_i, AP_i$ full rank (serious breakdown). The new algorithm inherits this problem, returning a division by zero if Ω_i or Π_i is singular. For $s = 1$ breakdown in Lanczos methods was analyzed by Gutknecht [16, 15]; the first work focusing on a look-ahead approach for Lanczos methods was that of Parlett, Taylor, and Liu [25], while further developments can be found in [11, 10, 3, 4, 2]. Determining a look-ahead strategy for the general case $s \geq 1$ would be very important for the robustness of QMR-MBCG. Although we will not address the general look-ahead problem, in section 5.1 we will describe a technique that improves the conditioning of Ω_i, Π_i by simply permuting some columns of the A-biorthogonalized matrices, thus exploiting the block form of the algorithm for recapturing the correct obliqueness between the spaces generated by the blocks. We will be referring to this modification of the code as *in-block* safeguard strategy.

Another source of failure is the loss of rank of the iteration matrices. In [23], O'Leary briefly suggested deflating the redundant columns while continuing to update the corresponding systems using the remaining iterates. A more detailed discussion on loss of rank was presented by Nikishin and Yeremin in [22]; see also the more

recent contribution [28]. In the symmetric case [22], it was observed that loss of rank is due to the exact solution of some of the systems of the modified problem $A\hat{X} = \hat{B}$, where $\hat{X} = X\delta$, $\hat{B} = B\delta$ for some nonsingular matrix δ. This idea can be generalized to the nonsymmetric case, observing that rank deficiency can occur in QMR-MBCG for any of the iterates Q_i and $\tilde{Q}_i$ or, equivalently, for S_i and $\tilde{S}_i$. Suppose first that $\sigma \equiv \text{rank}(Q_i) < s$ and $\text{rank}(\tilde{Q}_i) = s$. The condition $\sigma < s$ implies that $\text{span}\{Q_0, \ldots, Q_i\}$ contains the exact solution of $s - \sigma$ systems of $A\hat{X} = \hat{B}$, where here δ is such that $Q_i\delta = [Q_*, 0_{n,s-\sigma}]$. Matrix δ can be determined, for instance, by using a modification of the Gram–Schmidt procedure. It follows that $AX^* = B^*$ with $X_i\delta = [X_*, X^*]$, $B\delta = [B_*, B^*]$ [22]. Thus matrix X^* can be deflated and the process restarted for the remaining recurrence vectors with the smaller blocks X_*, B_* until convergence or the occurrence of a new loss of rank. Since we are interested in solving $AX = B$, we need to save all the δ's for recovering the final approximate solution of X.

If $\text{rank}(\tilde{Q}_i) < s$ and $\text{rank}(Q_i) = s$, an exact solution of a modification of $A^T\tilde{X} = \tilde{B}$ has been detected, where $\tilde{B}, \tilde{X}_0$ are such that $\tilde{R}_0 = \tilde{B} - A^T\tilde{X}_0$. Unfortunately, in general this is not a welcome event, since no exact solutions for $AX = B$ have been found. One can either restart the algorithm with a different choice of $\tilde{R}_0$ or eliminate the corresponding columns in all blocks, for which a better approximate solution can be determined separately. We also note that, in the single right-hand side case, restarting has been used to cure stagnation in the approximation process; see, for instance, [17].

5.1. Near-breakdown and safeguard strategy. We next analyze the problems of breakdown and near-breakdown in the Gram–Schmidt procedures MGS and MGS.A.

Let $Q = [q_1, \ldots, q_s]$ and $\tilde{Q} = [\tilde{q}_1, \ldots, \tilde{q}_s]$ and assume that the biorthogonalization is done in place, that is, at step k the first k columns of $Q, \tilde{Q}$ are biorthonormal. Consider the case where the process fails at step k; that is, $[\tilde{q}_1, \ldots, \tilde{q}_{k-1}]^T[q_1, \ldots, q_{k-1}]$ is nonsingular and diagonal and $[\tilde{q}_1, \ldots, \tilde{q}_k]^T[q_1, \ldots, q_k]$ is singular. This is equivalent to a zero diagonal element in matrix Ω of MGS.[2] In particular, this means that

$$q_k \perp \text{span}\{\tilde{q}_1, \ldots, \tilde{q}_k\} \quad \text{or} \quad \tilde{q}_k \perp \text{span}\{q_1, \ldots, q_k\}. \tag{5.1}$$

Without loss of generality, we can suppose that $q_k \perp \text{span}\{\tilde{q}_1, \ldots, \tilde{q}_k\}$. It was shown in [25] that if $s = n$ and $Q, \tilde{Q}$ have full rank, there exists $\bar{k} \in \{k+1, \ldots, s\}$ such that

$$q_k \not\perp \text{span}\{\tilde{q}_1, \ldots, \tilde{q}_k, \tilde{q}_{\bar{k}}\}. \tag{5.2}$$

In our application we assume $s \ll n$; therefore, such $\bar{k}$ is not assured to exist. If, however, we succeed in finding $\bar{k}$, we can permute columns k and $\bar{k}$ in $Q, \tilde{Q}$ and continue the process with the new vectors $q_{\bar{k}}, \tilde{q}_{\bar{k}}$ at position k. This allows us to avoid breakdown with vectors $q_k, \tilde{q}_k$. Note that the idea of permuting the columns stems from the look-ahead technique introduced in [25] for the single right-hand side Lanczos algorithm, although here it is used in the Gram–Schmidt biorthogonalization procedure.

In practice, $\bar{k}$ is found using a linear search starting at position $k+1$. This means that we anticipate the orthogonalization of all successive columns of $Q, \tilde{Q}$ with respect

[2] We restrict our discussion to MGS; however, it remains valid when applied to MGS.A.

to the first $k-1$ vectors until we find the first $\bar{k}$ satisfying (5.2). However, this work is not wasted, and the intermediate quantities can be saved for later use. Therefore, the computed vectors replace the original vectors at position $k+1,\dots,\bar{k}$, while the new biorthogonality coefficients can be allocated at the corresponding position of the triangular matrices $\xi, \tilde{\xi}$. At the following steps, the orthogonalization of each column q_l, for $l > k$, will be done only with respect to vectors $\tilde{q}_k,\dots,\tilde{q}_{l-1}$.

At the end of the process we need to resume the original order of the columns of the biorthogonalized matrices. If $\mathcal{E}$ is the resulting permutation matrix and X is the original right block of vectors, $X\mathcal{E} = (Q\mathcal{E}^T)(\mathcal{E}\xi)$. Thus it is sufficient to set

$$Q \leftarrow Q\mathcal{E}^T, \quad \xi \leftarrow \mathcal{E}\xi\mathcal{E}^T, \quad \text{and} \quad \Omega \leftarrow \mathcal{E}\Omega\mathcal{E}^T.$$

Matrices $\tilde{Q}$ and $\tilde{\xi}$ are set analogously. We note that the permutation could also be carried out in an unsymmetric order, that is, by permuting only one of the two matrices $Q, \tilde{Q}$. This technique would allow us to address either of the two conditions in (5.1).

In finite arithmetic, the case when Ω is almost singular (near-breakdown) is of interest and corresponds to

$$\Omega_{k,k} \equiv \tilde{q}_k^T q_k < \epsilon \tag{5.3}$$

for some $k \in \{1,\dots,s\}$ with ϵ a chosen threshold. The test (5.3) replaces (5.1) in the practical implementation. If the linear search is unsuccessful and all subsequent pairs of vectors satisfy (5.3), the permutation is carried out with those vectors whose inner product gives the closest value to the threshold. However, since $\bar{k} \notin \{k+1,\dots,s\}$, the next iterations may suffer from ill conditioning if ϵ is small. We remark that the selection of the threshold is crucial for achieving the best performance of the method since $\|\Omega^{-1}\|$ affects the magnitude of the recurrence coefficients. This suggests that a strict value of ϵ could be advantageous in several cases (cf. section 6.1).

We finally remark that an equivalent procedure can be applied if one allows Ω to be block diagonal instead of diagonal. This was called by Parlett the "extended" Gram–Schmidt procedure [24]. If near-breakdown is detected at step k, biorthogonality with respect to vectors $q_k, q_{k+1},\dots$ and $\tilde{q}_k, \tilde{q}_{k+1},\dots$ is relaxed until an index $\bar{k} \in \{k+1,\dots,s\}$ is found such that the matrix $[\tilde{q}_k, \tilde{q}_{k+1},\dots,\tilde{q}_{\bar{k}}]^T[q_k, q_{k+1},\dots,q_{\bar{k}}]$ has smallest singular value greater than the threshold. In our experiments, however, we have used the symmetric permutation approach, which showed the best overall performance.

6. Numerical results. The aim of this section is to compare the new method QMR-MBCG with the original BBICG method and the single right-hand side solver coupled two-term QMR (CQMR) proposed in [12].

Memory requirements strongly depend on s for all algorithms. In particular, for QMR-MBCG $8ns$ locations are needed for the recurrence iterates; parameters $\xi_i, \tilde{\xi}_i, \tau_i, \tilde{\tau}_i$, and $\eta_i, \tilde{\eta}_{i+1}$ are all triangular matrices and thus require $3s(s+1)$ locations; finally, $3s$ locations are needed for Ω_i, Ω_{i+1}, and Π_i, for a total of $8ns + 3s^2 + 6s$ of minimum memory allocation, compared to the $5ns + \frac{9}{2}s^2 + \frac{1}{2}s$ of BBICG and $6n$ of CQMR for each system. Additional memory may be needed depending on the particular implementation. Note that the requirements of QMR-MBCG for $n \times s$ matrices could drop to 6 by not storing matrices $L_i, \tilde{L}_i$; however, this would yield a slightly higher computation cost per iteration.

The major computational cost per iteration of each method is shown in Table 6.1 in floating point operations; all methods also require $2s$ multiplications with A and

TABLE 6.1
Major computational cost per iteration with s right-hand sides and $n = \dim(A)$. All methods also require $2s$ multiplications with A and A^T per step.

method	CQMR	BBICG	QMR-MBCG
flops	$22ns$	$16ns^2 + 8ns + \frac{44}{3}s^3$	$23ns^2 + 23ns + 6s^3$

A^T per step. We have considered that block DAXPY's correspond to $2ns^2 + ns$ flops and that the solution of triangular systems with s right-hand sides corresponds to s^3 flops. Standard MGS for orthogonalizing a matrix costs $2ns^2 + 3ns$ flops [14]. The cost of CQMR is computed for all systems. The cost of evaluating the norm of the true residual at each step has not been included in the table. The large number of flops for QMR-MBCG is due to the double biorthogonalization and the quasi minimization performed at each step. On the other hand, computation with $s \times s$ matrices decreases considerably.

In general, block methods show a very high computational cost per iteration. However, they are expected to converge in fewer iterations than using the single right-hand side solver; see [29] for a quantitative analysis of the global cost of short-term recurrence block methods.

6.1. Tests. The first set of experiments is devoted to the analysis of the robustness of QMR-MBCG, whereas in the second set we will deal with timings performance. BBICG was implemented as described in [23] with orthogonalization of the direction matrices $P_i, \tilde{P}_i$. QMR-MBCG was applied with in-block safeguard strategy on both routines MGS and MGS.A. In all tests we used a zero starting guess, so that $R_0 = B$. Unless otherwise stated, we set $\tilde{R}_0 = R_0$. Other selections are possible [17]; however, the chosen $\tilde{R}_0$ did not cause failure of the process since the new version of the algorithm converges in all cases. A system was considered sufficiently well approximated when the norm of the true relative residual of each system was less than $\varepsilon = 10^{-6}$.

Set 1. The following experiments were run with Matlab [21] on an SGI Workstation with round-off unit $eps \approx 2.22 \cdot 10^{-16}$. We considered the matrix A resulting from the discretization by central finite differences of the elliptic operator [12]

$$L(u) = -\left(e^{-xy}u_x\right)_x - \left(e^{xy}u_y\right)_y + 20(x+y)u_x + 20((x+y)u)_x + \frac{1}{1+x+y}u \tag{6.1}$$

on the unit square grid with Dirichlet boundary conditions. The number of grid points was chosen so that the final dimension of A was $n = 2500$. We considered B with random values and $s = 4$. Unless otherwise specified, we used near-breakdown threshold $\epsilon = eps$.

Let $\mathcal{S} = [S_0, S_1, \ldots]$, $\mathcal{P} = [P_0, P_1, \ldots]$, and $\tilde{\mathcal{S}}$, $\tilde{\mathcal{P}}$ accordingly defined. In Figure 6.1 the sparsity pattern of $\tilde{\mathcal{P}}^T A\mathcal{P}$ and $\tilde{\mathcal{S}}^T A\mathcal{S}$ after the first 20 iterations[3] is reported, where the dots correspond to elements with absolute value larger than 10^{-9}. In exact arithmetic, $\tilde{\mathcal{S}}^T A\mathcal{S}$ and $\tilde{\mathcal{P}}^T A\mathcal{P}$ are diagonal and block diagonal matrices, respectively, whereas Figure 6.1 shows that this property is not preserved in finite precision arithmetic. The loss of orthogonality is not surprising, although we observe better orthogonality properties of the new iterates $\{S_i\}, \{\tilde{S}_i\}$. It was already pointed out for the symmetric block Lanczos algorithm that loss of orthogonality is strictly related to the growth of the coefficients norm [20]. This is indeed true in the nonsymmetric case,

[3] The small number of iterations is only due to printer constraints; the sparsity pattern remains consistent for a higher number of iterations.

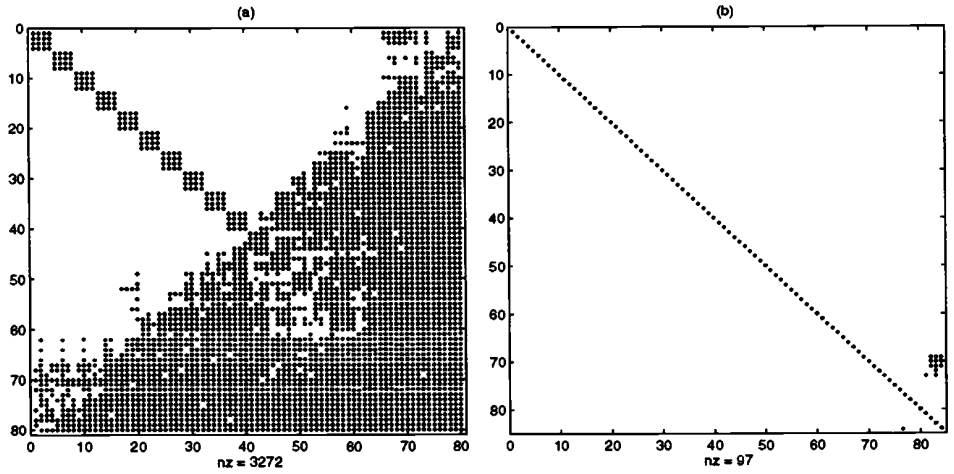


FIG. 6.1. *Test for A-biorthogonality for matrix A from* (6.1) *after* 20 *iterations.* $n = 2500$ *and four right-hand sides.* (a): $|\tilde{\mathcal{P}}^T A\mathcal{P}| > 10^{-9}$ *for BBICG*; (b): $|\tilde{\mathcal{S}}^T A\mathcal{S}| > 10^{-9}$ *for QMR-MBCG.*

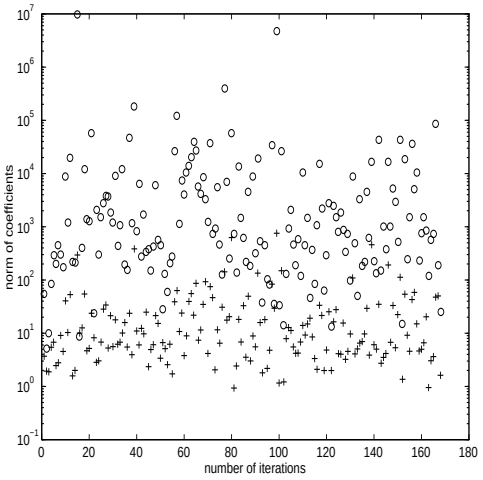


FIG. 6.2. *Norm of coefficients* α_i *("o") and* σ_i *("+") of BBICG and QMR-MBCG, respectively.*

and the same phenomenon can be observed in the BBICG algorithm. However, the coefficients of the new version of the algorithm are characterized by smaller norms. Figure 6.2 reports the values of $\|\alpha_i\|$ from formula (2.4) for BBICG ("o") and the values of $\|\sigma_i\|$ from formula (4.3) for QMR-MBCG ("+") for 160 iterations. It can be easily shown from the derivation of the recurrence terms of QMR-MBCG that the two sets of coefficients are linked by the relation $\alpha_i = (\xi_i \dots \xi_0)^{-1}\sigma_i(\xi_i \dots \xi_0)$, which indicates that α_i may be affected by the condition number of the biorthogonalization factors $\xi_i \dots \xi_0$. A similar relation holds for $\tilde{\alpha}_i$ and $\tilde{\sigma}_i$.

In Figure 6.3(a), the convergence history of all the algorithms is shown. The plateau during the first phase of QMR-MBCG corresponds to a diverging portion of the BBICG residual curve, and it seems to be a pattern of the QMR procedure [6]. It also appears that BBICG is not able to recover from an early divergence state.

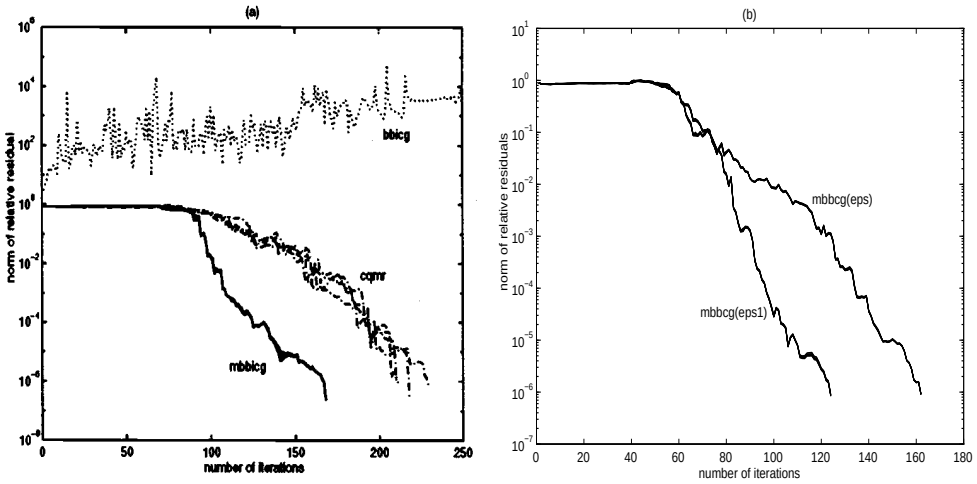


FIG. 6.3. (a): *Performance of methods in number of iterations with matrix from* (6.1)*,* $n = 2500$*,* $s = 4$; (b): *influence of the near-breakdown threshold on QMR-MBCG with* $\epsilon = eps$ *and* $\epsilon \equiv eps1 = eps^{1/5}$.

Note, however, that a possibly erratic convergence curve does not in general prevent the method from converging [6]. For this experiment, the lack of convergence of BBICG is due to the loss of biorthogonality of the basis iterates. CQMR applied on each system converges in more iterations than QMR-MBCG, although the cost is not comparable (cf. Table 6.1).

Figure 6.3(b) shows the behavior of QMR-MBCG applied on A with $s = 10$ for $\epsilon = eps$ and $\epsilon = eps1$ with $eps1 = eps^{\frac{1}{5}}$. We note that the selection of a different threshold considerably improves the convergence history. The first permutation is done after four iterations. Note that if ϵ is taken to be large, the safeguard procedure in practice determines matrices Ω_i, Π_i with largest smallest singular values, benefiting the magnitude of the recurrence coefficients.

Set 2. These experiments were performed with matrices from the Harwell–Boeing collection [8]. The density of the matrix or the cost of preconditioning is such that block methods become of interest. BBICG was implemented so as to eliminate the converged systems while continuing the recurrence with the systems that necessitate additional iterations for convergence. No restarting was applied in order to test the performance of the original method. For a fair comparison with BBICG, deflation of converged systems for QMR-MBCG was not implemented for these examples. Indeed, due to the quasi-minimization process, elimination of systems would be carried out by restarting the algorithm with the remaining columns of the residual matrix and thus including new parameters in the computation. The safeguard threshold was $\epsilon = eps$, and a maximum of 200 iterations was allowed.

All tests were carried out on an Alliant FX/2800 (running Concentrix 2.2) using one processor. The $s \leq 20$ right-hand sides were chosen as random numbers with uniform distribution in $[0, 1]$ (fortran function `drand`). Computation was done in real double precision arithmetic and A was stored in sparse row format [27].

In Table 6.2 the total CPU time (fortran function `dtime`) for the convergence of each method is shown. For each matrix the dimension (n) and the number of nonzero elements (nnz) are listed. When used, the ILU preconditioner is also marked, where

TABLE 6.2

Total CPU time, in seconds, for methods to convergence with matrices from the Harwell–Boeing collection. "" stands for convergence affected by round-off; "-" stands for stagnation.*

	s	4	8	12	16	20
matrix	method					
PSMIGR3	CQMR	243	490	731	962	1210
$n = 3140$	BBICG	269	-	-	-	-
$nnz = 543162$	QMR-MBCG	277	371	308	373	433
ORSREG1(ILU(0))	CQMR	28	56	83	112	141
$n = 2205$	BBICG	-	-	-	-	-
$nnz = 14133$	QMR-MBCG	44	77	81	102	121
SAYLR4(ILU(5))	CQMR	103	206	342	478	583
$n = 3564$	BBICG	-	-	-	-	-
$nnz = 22316$	QMR-MBCG	105	193	245	336	417
SHERMAN3(ILU(3))	CQMR	60	121	160	254	310
$n = 5005$	BBICG	61*	-	-	-	152*
$nnz = 20033$	QMR-MBCG	48	67	91	117	142
SHERMAN5(ILU(0))	CQMR	26	49	74	96	124
$n = 3312$	BBICG	23	-	40	53	82
$nnz = 20793$	QMR-MBCG	27	45	51	68	94

ILU(θ) carries out an incomplete (with θ additional fill-in elements) LU decomposition of A.

We first note the disappointing behavior of BBICG, which stagnates (symbol "-" in Table 6.2) for several choices of s. Note also that for SHERMAN3 the method converges, but round-off delays the convergence. We need to add that for this test matrix B was set to the principal part of the identity matrix and $\tilde{R}_0$ was chosen to have random components. In successful runs, however, BBICG is less expensive than QMR-MBCG. We also remark that loss of rank of the iteration matrices was never encountered in these examples, whereas matrix β_i was found to be numerically singular ($\sigma_{\min}(\beta_i) < 10^{-13}$) in all failures. This was detected, in later runs, by computing both the singular value decomposition of β_i and the QR decomposition of $R_i, \tilde{R}_i$ at each iteration. The new block version stabilizes the recurrence, yielding an overall good performance, compared to the single right-hand side solver. We would also like to point out the improvement in CPU time per right-hand side as s increases; this shows that, for the examples reported, sharing of information does lead to a good performance of block methods.

6.2. Some applications. Both methods BBICG and QMR-MBCG generate the block Krylov subspaces span$\{R_0, AR_0, A^2R_0, \ldots\}$ and span$\{\tilde{R}_0, A^T\tilde{R}_0, (A^T)^2\tilde{R}_0, \ldots\}$. If $\tilde{R}_0$ corresponds to the residual of $A^T\tilde{X} = \tilde{B}$, for given $\tilde{X}_0, \tilde{B}$, both algorithms can determine an approximate solution of $A^T\tilde{X} = \tilde{B}$ at a very low additional cost.

The dependence on an auxiliary system is usually considered a disadvantage [17]. However, in some application problems where both A and A^T can be exploited the biorthogonal formulation is a successful alternative, as is shown in the examples below. The reliability of QMR-MBCG over the original block algorithm makes such an approach even more attractive.

The first case concerns the computation of the product $\sigma = C^TA^{-1}B \in \mathbb{R}^{s\times s}$, of importance, for instance, in system control theory and domain decomposition methods; see, for instance, [13]. Note that σ can also be computed as $\sigma = (B^TA^{-T}C)^T$. Thus, it may be convenient to compute both approximants X_m and $\tilde{X}_m$ of $A^{-1}B$ and $A^{-T}C$, respectively, and stop when any of the residuals $\tilde{R}_m = C - A^T\tilde{X}_m$, $R_m = B - AX_m$ is sufficiently small.

The second case corresponds to the computation of an $s \times s$ section of the inverse of a nonsingular matrix A. Let $\mathcal{A} = A^{-1}$ and $\mathcal{A} = (\mathcal{A}_{i,j})$. For instance, $\mathcal{A}_{1,1}$ can be approximated by the first s rows of the solution of $AX = I_{n,s}$ or, equivalently, by the transpose of the first s rows of the solution of $A^T\tilde{X} = I_{n,s}$, since $X_{1:s,1:s} = \mathcal{A}_{1,1} = \tilde{X}^T_{1:s,1:s}$. The procedure can be stopped as soon as one of the approximate solutions reaches the requested tolerance. As an example of such eventuality, consider matrix A, arising in the discretization by finite differences of the operator

$$L(u) = -\Delta u + 50\,(xu_x)_x + 50\,(yu_y)_y$$

with $n = 2500$ and $s = 6$; here we used $\epsilon = eps$. In Figure 6.4(a) we have plotted the behavior of the true relative residual norm (with Matlab) for both systems $AX = I_{n,s}$ (symbol "–") and $A^T\tilde{X} = I_{n,s}$ (symbol ":") using BBICG. Analogously, Figure 6.4(b) shows the convergence history of the residuals of the two systems with A and A^T using QMR-MBCG. Using both algorithms, the transpose system converges faster than the system with A so that a rough approximation to $\mathcal{A}_{1,1}$ could be obtained in an efficient manner.

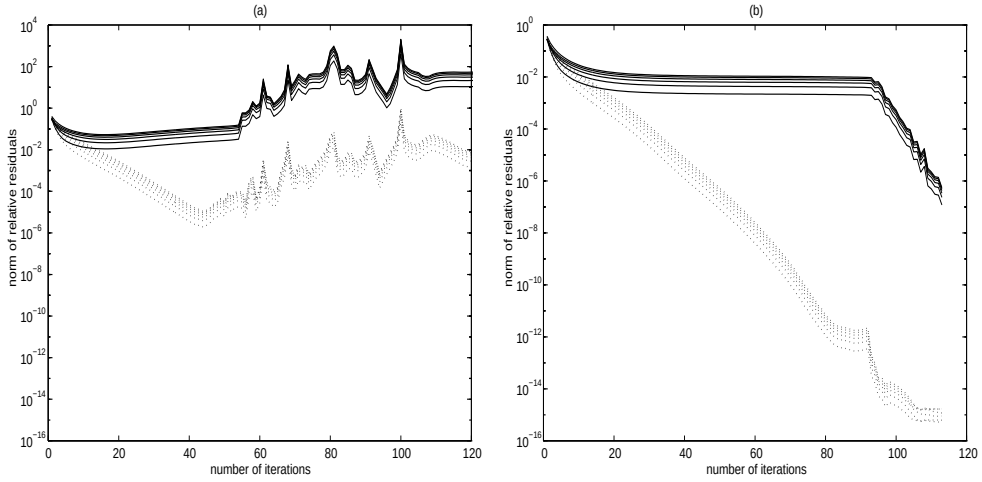


FIG. 6.4. *Convergence history with matrix A from* (6.1) *and* $B = I_{n,s}$ *with* $n = 2500$ *and* $s = 6$. (a): *BBICG*; (b): *QMR-MBCG. In both cases, "–" refers to the norm of* $R_m = B - AX_m$ *and ":" refers to that of* $\tilde{R}_m = B - A^T\tilde{X}_m$.

On the other hand, this example indicates that the recurrence on the transpose system influences the computation. The residuals $R_m, \tilde{R}_m$ are directly involved in the recurrence of BBICG (2.2); this means that the magnitude of their elements may affect the precision of the computation itself. This may justify the deterioration of convergence in BBICG, whereas QMR-MBCG can successfully terminate since the elements of the new sets of iterates maintain the same order of magnitude.

7. Conclusions. We summarize our conclusions as follows.

(i) The vectorwise A-biorthogonalization of the iterates yields a more robust method than BBICG.

(ii) Recurrences that do not involve the residuals are less sensitive to the instability due to the magnitude of the elements.

(iii) The in-block safeguard strategy procedure introduced may improve the convergence behavior of the algorithm.

(iv) QMR-MBCG can be very useful in some application problems involving both A and A^T, since it outperforms the single right-hand side solver CQMR in several cases.

The new version of block BICG considerably ameliorates the original method. However, we feel that a global look-ahead procedure accompanied by a simple deflation strategy would be very desirable for the implementation of a robust, short-term recurrence algorithm for solving systems with multiple right-hand sides.

Acknowledgments. We are extremely grateful to Prof. D. O'Leary and Prof. B. Parlett for enlightening discussions on BBICG and orthogonalization, and to Prof. E. Gallopoulos for a continuous exchange of opinions on the topic. The suggestions of the anonymous referees helped improve the presentation of the paper considerably. We would also like to thank CSRD for allowing the use of the facilities of the center.

REFERENCES

[1] M. Arioli, I. Duff, D. Ruiz, and M. Sadkane, *Techniques for accelerating the block Cimmino method*, in Proc. Fifth SIAM Conf. Parallel Processing for Scientific Computing, J. J. Dongarra, K. Kennedy, P. Messina, D. C. Sorensen, and R. G. Voigt, eds., SIAM, Philadelphia, PA, 1992, pp. 98–104.

[2] R. E. Bank and T. F. Chan, *An analysis of the composite step biconjugate gradient method*, Numer. Math., 66 (1993), pp. 293–319.

[3] D. L. Boley and G. H. Golub, *The nonsymmetric Lanczos algorithm and controllability*, Systems Control Lett., 16 (1991), pp. 97–105.

[4] C. Brezinski, M. R. Zaglia, and H. Sadok, *A breakdown-free Lanczos type algorithm for solving linear systems*, Numer. Math., 63 (1992), pp. 29–38.

[5] D. Choudhury and R. A. Horn, *An Analog of the Gram–Schmidt Algorithm for Complex Bilinear Forms and Diagonalization of Complex Symmetric Matrices*, Tech. report 454, The Johns Hopkins University, Baltimore, MD, 1986.

[6] J. K. Cullum, *Peaks and Plateaus in Lanczos Methods for Solving Nonsymmetric Systems of Equations $Ax = b$*, Tech. report RC 18084, IBM T. J. Watson Research Center, 1992.

[7] I. S. Duff, A. M. Erisman, and J. K. Reid, *Direct Methods for Sparse Matrices*, Clarendon Press, Oxford, 1989.

[8] I. S. Duff, R. G. Grimes, and J. G. Lewis, *User's Guide for the Harwell–Boeing Sparse Matrix Collection* (*Release* I), Tech. report TR/PA/92/86, CERFACS, Toulouse Cedex, France, 1992.

[9] R. W. Freund, *Transpose-free quasi-minimal residual methods for non-Hermitian linear systems*, in Recent Advances in Iterative Methods, IMA Volumes in Mathematics and Its Applications 60, G. H. Golub, A. Greenbaum, and M. Luskin, eds., Springer-Verlag, New York, 1994, pp. 69–94.

[10] R. W. Freund, M. H. Gutknecht, and N. M. Nachtigal, *An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices*, SIAM J. Sci. Comput., 14 (1993), pp. 137–158.

[11] R. W. Freund and N. M. Nachtigal, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.

[12] R. W. Freund and N. M. Nachtigal, *An implementation of the QMR method based on coupled two-term recurrences*, SIAM J. Sci. Comput., 15 (1994), pp. 313–337.

[13] I. Gohberg, P. Lancaster, and L. Rodman, *Invariant Subspaces of Matrices with Applications*, John Wiley, New York, 1986.

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

[15] M. H. Gutknecht, *A Completed Theory of the Unsymmetric Lanczos Process and Related Algorithms*. II, Tech. report 90.16, IPS, ETH-Zentrum, 1990.

[16] M. H. Gutknecht, *A completed theory of the unsymmetric Lanczos process and related algorithms*. I, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 594–639.

[17] W. D. Joubert, *Lanczos methods for the solution of nonsymmetric systems of linear equations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 926–943.

[18] S. Kharchenko, P. Kolesnikov, A. Nikishin, A. Yeremin, M. Heroux, and Q. Seikh, *Iterative solution methods on the Cray YMP/C*90. *Part* II: *Dense linear systems*, presented

at 1993 Simulation Conference: High Performance Computing Symposium, Washington D.C.

[19] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Nat. Bur. Standards, 45 (1950), pp. 255–282.

[20] J. Lewis, *Algorithms for sparse matrix eigenvalue problems*, Tech. report STAN-CS-77-595, Department of Computer Science, Stanford University, Stanford, CA, 1977.

[21] The MathWorks, Inc., *MATLAB User's Guide*, Natick, MA, 1993.

[22] A. A. Nikishin and A. Y. Yeremin, *Variable block CG algorithms for solving large sparse symmetric positive definite linear systems on parallel computers,* I: *General iterative scheme*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1135–1153.

[23] D. P. O'Leary, *The block conjugate gradient algorithm and related methods*, Linear Algebra Appl., 29 (1980), pp. 293–322.

[24] B. N. Parlett, *Reduction to triangular form and minimal realizations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 567–593.

[25] B. N. Parlett, D. R. Taylor, and Z. A. Liu, *A look-ahead Lanczos algorithm for unsymmetric matrices*, Math. Comp., 44 (1985), pp. 105–124.

[26] A. Ruhe, *Implementation aspects of band Lanczos algorithms for computation of eigenvalues of large sparse symmetric matrices*, Math. Comp., 33 (1979), pp. 680–687.

[27] Y. Saad, *SPARSKIT: A Basic Tool Kit for Sparse Matrix Computation*, Tech. report 1029, Center for Supercomputing Research and Development, University of Illinois at Urbana-Champaign, 1990.

[28] H. D. Simon and A. Yeremin, *A new approach to construction of efficient iterative schemes for massively parallel algorithms: Variable block CG and BiCG methods and variable block Arnoldi procedure*, in Proc. Sixth SIAM Conf. Parallel Processing for Scientific Computing, R. F. Sincovec, D. E. Keyes, M. R. Leuze, L. R. Petzold, and D. A. Reed, eds., Philadelphia, PA, 1993, pp. 57–60.

[29] V. Simoncini and E. Gallopoulos, *Iterative Methods for Complex Symmetric Systems with Multiple Right-Hand Sides*, Tech. report 1322, Center for Supercomputing Research and Development, University of Illinois at Urbana-Champaign, 1993.

[30] V. Simoncini and E. Gallopoulos, *An iterative method for nonsymmetric systems with multiple right-hand sides*, SIAM J. Sci. Comput., 16 (1995), pp. 917–933.

[31] V. Simoncini and E. Gallopoulos, *Convergence properties of block GMRES and matrix polynomials*, Linear Algebra Appl., 247 (1996), pp. 97–119.

[32] B. Vital, *Etude de quelques méthodes de résolution de problèmes linéaires de grande taille sur multiprocesseur*, Ph.D. thesis, Université de Rennes I, Rennes, 1990.

[33] L. Zhou and H. F. Walker, *Residual smoothing techniques for iterative methods*, SIAM J. Sci. Comput., 15 (1994), pp. 297–312.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, pp. 435–449, April 1997

013

ANALYSIS OF AUGMENTED KRYLOV SUBSPACE METHODS*

YOUSEF SAAD†

Abstract. Residual norm estimates are derived for a general class of methods based on projection techniques on subspaces of the form $K_m + \mathcal{W}$, where K_m is the standard Krylov subspace associated with the original linear system and $\mathcal{W}$ is some other subspace. These "augmented Krylov subspace methods" include eigenvalue deflation techniques as well as block-Krylov methods. Residual bounds are established which suggest a convergence rate similar to one obtained by removing the components of the initial residual vector associated with the eigenvalues closest to zero. Both the symmetric and nonsymmetric cases are analyzed.

Key words. Krylov methods, deflated iterations, block-GMRES

AMS subject classification. 65F

PII. S0895479895294289

1. Introduction.

It has recently been observed that significant improvements in convergence rates can be achieved from Krylov subspace methods by enriching these subspaces in a number of different ways; see, e.g., [2, 4, 8, 9]. One of the simplest ideas employed is to add to the Krylov subspace some approximation to an invariant subspace associated with a few of the lowest eigenvalues. A projection process on this augmented subspace is then carried out. An older technique is to augment the original subspace with other Krylov subspaces, typically with the same matrix and randomly generated right-hand sides. This gives rise to the class of block-Krylov and successive right-hand side methods which have recently seen a resurgence of interest [14, 11, 1, 6, 5]. Results of experiments obtained from these alternatives indicate that the improvement in convergence over standard Krylov subspaces of the same dimension can sometimes be substantial. This is especially true when the convergence of the original scheme is hampered by a small number of eigenvalues near zero; see e.g., [2, 9].

In this paper we take a theoretical look at this general class of "augmented Krylov methods." In short, an augmented Krylov method for solving the linear system

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

is any projection method in which the subspace of projection is of the form

$$K = K_m + \mathcal{W},$$

where K_m is the standard Krylov subspace

$$K_m = \text{span}\{r_0, Ar_0, \ldots, A^{m-1}r_0\}$$

with $r_0 = b - Ax_0$, the vector x_0 being an arbitrary initial guess to the above linear system. Thus, the usual Krylov subspace K_m, which we sometimes call the primary

*Received by the editors November 3, 1995; accepted for publication (in revised form) by R. Freund June 5, 1996. This work was supported in part by NASA grant NAG2-904 and in part by NSF grant CCR-9214116. This research was conducted while the author was on leave at the University of California at Los Angeles.

http://www.siam.org/journals/simax/18-2/29428.html

†University of Minnesota, Department of Computer Science, Minneapolis, MN 55455 (saad@cs.umn.edu).

subspace, is augmented by another subspace $\mathcal{W}$. The intuitive rationale for these methods is that K_m cannot always capture all the "frequencies" of A, so it may become necessary to include explicitly those components which cause the method to slow down. There are many possible ways in which to choose the subspace $\mathcal{W}$ following this intuitive idea. In deflation techniques [9, 2], $\mathcal{W}$ is an approximate invariant subspace typically associated with the smallest eigenvalues and obtained as a by-product of earlier projection steps. In block-Krylov techniques, $\mathcal{W}$ consists of the sum (in the linear algebra sense) of a few other Krylov subspaces generated with the same matrix A, but different initial residuals.

We now give a brief background and define some terminology. In what follows, $\mathbb{P}_k$ denotes the space of polynomials of degree not exceeding k, while $\mathbb{P}_k^*$ is the space of polynomials p of degree $\leq k$ normalized so that $p(0) = 1$. An invariant subspace is any subspace X of $\mathbb{C}^n$ such that AX is included in X. If $W = [w_1, \ldots, w_p]$ is a basis of X then X is invariant iff there is a $p \times p$ matrix G such that $AW = WG$. In this paper we often use projections of vectors onto invariant subspaces. This can be done in several ways. Two important options are to use either orthogonal projectors onto the invariant subspace or spectral projectors. A spectral projector is best defined through the Jordan canonical form. The Jordan canonical form decomposes the subspace $\mathbb{C}^n$ into the direct sum

$$\mathbb{C}^n = X_1 \oplus X_2 \oplus \cdots \oplus X_l,$$

in which each X_i is the invariant subspace associated with a distinct eigenvalue. This direct sum defines canonically a set of l projectors. Each of these projectors maps an arbitrary vector x into its component x_i in the above decomposition. A spectral projector is the sum of any number of these canonical projectors.

Two types of methods are often used to compute an approximate solution from a given subspace. An orthogonal projection method, or orthogonal residual (orth-res) method, extracts an approximation solution of the form $x = x_0 + \delta$, where δ is in K, by imposing the orthogonality constraint $b - Ax \perp K$. A minimal residual (min-res) approach computes an approximation of the same form but extracts the approximation by imposing the optimality condition that $\|b - Ax\|_2$ be minimal. This second condition is mathematically equivalent to the orthogonality condition that $b - Ax \perp AK$.

2. Augmented Krylov methods and flexible GMRES (fGMRES). To obtain an orthogonal basis of an augmented Krylov subspace, a slight modification of the standard Arnoldi algorithm is needed. Assume that we have a subspace spanned by $m+p$ vectors. Specifically, the first m of these vectors are standard Krylov vectors $v_1, \ldots, v_m$, and the last ones, denoted by $w_1, \ldots, w_p$, form a basis of the additional subspace $\mathcal{W}$. Then at step $m+1$ we introduce the first basis vector w_1 of $\mathcal{W}$, multiply it by A as in the Arnoldi process, and orthogonalize the result against all previous vectors. We then similarly introduce the next basis vector to the subspace and repeat this process. The algorithm is as follows.

ALGORITHM 2.1 (*augmented Arnoldi-modified Gram–Schmidt*).

1. *Choose a vector* v_1 *of norm* 1.
2. *For* $j = 1, 2, \ldots, m + p$ *Do:*
3. *If* $j \leq m$ *then* $w := Av_j$, *Else* $w := Aw_{j-m}$
4. *For* $i = 1, \ldots, j$ *do:*
5. $h_{ij} = (w, v_i)$
6. $w := w - h_{ij} v_i$

7. *EndDo*
9. $h_{j+1,j} = \|w\|_2$. *If* $h_{j+1,j} = 0$ *then Stop.*
10. $v_{j+1} = w/h_{j+1,j}$
11. *EndDo*

We can think of many possible variations to the above basic scheme. For example, the input vectors w_i can themselves be the Krylov vectors of some iterative procedure for solving $Aw = v_{m+1}$. We can also generate another Krylov sequence starting with an arbitrary vector w_1 and append the resulting vectors $w_2, \ldots, w_p$ to the subspace. Some of these variations are explored in [2].

The above algorithm is a trivial extension of the modified Arnoldi process used in the fGMRES algorithm [12]. Its result is that the vectors $v_1, \ldots, v_{m+p+1}$ form an orthonormal set of vectors. A number of immediate properties can be established. First, the vectors produced by the algorithm satisfy the relation

$$AZ_{m+p} = V_{m+p+1}\bar{H}_m,$$

in which

$$Z_{m+p} = [v_1, v_2, \ldots, v_m, w_1, w_2, \ldots, w_p], \qquad V_{m+p+1} = [v_1, v_2, \ldots, v_{m+p+1}],$$

and $\bar{H}_m$ is the $(m+p+1)\times(m+p)$ upper Hessenberg matrix whose nonzero elements h_{ij} are defined in the algorithm. To solve a linear system with an fGMRES-like approach, we only need to exploit the above relation and the orthogonality of the v_i's. Thus, if $\beta := \|r_0\|_2$ and we start the Arnoldi process with $v_1 := r_0/\beta$, then an approximate solution x from the affine space $x_0 + \text{span}\{Z_{m+p}\}$ can be written in the form $x_0 + Z_{m+p}y$ and its residual vector is given by

$$b - Ax = r_0 - AZ_{m+p}y = V_{m+p+1}[\beta e_1 - \bar{H}_m y].$$

Because of the orthogonality of the column vectors of V_{m+p+1}, the 2-norm of this residual vector can be minimized by solving the least-squares problem $\min_y \|\beta e_1 - \bar{H}_m y\|_2$.

Another important property is that if any vector w in $\mathcal{W}$ is the solution of an equation $Aw = v_i$ for any of the v_i's, $i \le m+1$, then, in general, the exact solution can be extracted from the whole subspace by an fGMRES procedure.

PROPOSITION 2.1. *If there exists a vector* w *in* $\mathcal{W}$ *such that* $Aw = v_{i+1}$ *for some* i, $1 \le i \le m$, *and if the matrix* H_i *is nonsingular then the affine space* $x_0 + K_m + \mathcal{W}$ *contains an exact solution to the linear system* $Ax = b$.

Proof. Assume that w is a vector in $\mathcal{W}$ such that $Aw = v_{i+1}$. Recall the standard relation [13]

$$AV_i = V_i H_i + h_{i+1,i} v_{i+1} e_i^T. \tag{2.1}$$

A solution among vectors of the form

$$x = x_0 + V_i y + \alpha w$$

will be constructed. For such vectors the residual $b - Ax$ is given by

$$r_0 - AV_i y - \alpha Aw = V_i\,(\beta e_1 - H_i y) - (h_{i+1,i} e_i^T y + \alpha) v_{i+1}.$$

If H_i is nonsingular, then y can be chosen so that the first term in the right-hand side vanishes. The scalar α can then be selected to be equal to $-h_{i+1,i} e_i^T y$ to make the second term equal to zero. ∎

In the situation of the proposition, fGMRES will compute the exact solution. This is because fGMRES extracts the (unique) approximate solution with minimum residual. In fact, any projection procedure onto the subspace $x_0 + K_m + \mathcal{W}$ will extract this exact solution because a solution with zero residual can be obtained from the subspace, and therefore the Galerkin condition will always be satisfied for this (exact) solution. Note that the proposition is also trivially true for $i = 0$, with the exception that we no longer need the assumption on H_i which does not exist. In addition, it can also be generalized to the situation where there is a vector w in $\mathcal{W}$ such that $Aw = v$ for some vector v in K_{m+1}.

Proposition 2.1 suggests that a good way to enrich the subspace K_m is to add to it vectors $w_1, \ldots, w_p$ that are approximate solutions of the linear system $Aw = v_i$ for $i \leq m+1$. These linear systems can be solved with a different preconditioner, for example, one which complements the initial one used for the primary linear system being solved. In effect, we can view this as a multirate approach. The Krylov subspace K_m is often unable to resolve components of the residual vector that are located in some subspace. Roughly speaking, much of the work in solving the linear system is already accomplished by the subspace K_m. The additional subspace will then fine-tune the current solution in the areas of the spectrum which are not well represented by K_m. In the simplest case, one can add solutions of linear systems $Aw = v_{m+1}$ by another iteration method such as a multistep SOR. An interesting idea which has been quite successful is to take $\mathcal{W}$ to be an approximate invariant subspace associated with small eigenvalues.

3. Augmenting with nearly invariant subspaces. In what follows we denote by x_0 the initial guess used in the augmented GMRES process for solving the linear system (1.1), by r_0 the associated initial residual $b - Ax_0$, and by K_m the Krylov subspace

$$K_m(A, r_0) = \text{span}\{r_0, Ar_0, \ldots, A^{m-1}r_0\}.$$

We make the assumption that there exists an invariant subspace which is close to $\mathcal{W}$ and analyze the behavior of the resulting augmented Krylov subspace algorithm. Our goal is to show a residual bound indicating faster convergence when the invariant subspace is very close to $\mathcal{W}$.

3.1. Basic results. We recall the following definition of the "gap" between subspaces. For details of this definition and some properties, see Kato [7] and Chatelin [3].

DEFINITION 3.1. *For any pair of subspaces of* $\mathbb{C}^n$ *define*

$$\delta(X, Y) = \max_{x \in X,\ x \neq 0} \min_{y \in Y} \frac{\|x - y\|_2}{\|x\|_2}. \tag{3.1}$$

Then the gap between the subspaces X *and* Y *is*

$$\Theta(X, Y) = \max\left[\delta(X, Y), \delta(Y, X)\right]. \tag{3.2}$$

Thus, $\delta(X, Y)$ represents the sine of the largest possible angle between vectors in X and their projections in Y. It is worth pointing out that $\delta(X, Y) = \|(I - P_Y)P_X\|_2$, in which P_X (resp., P_Y) is an orthogonal projector onto X (resp., Y). In fact, when the two subspaces X and Y are of the same dimension then [3, 7]

$$\Theta(X, Y) = \delta(X, Y) = \delta(Y, X) = \|P_X - P_Y\|_2.$$

In this case, $\Theta(X, Y)$ can be viewed as the sine of the angle between the two subspaces X and Y.

THEOREM 3.2. *Assume that a min-res projection method is applied to A using the augmented Krylov subspace*

$$K = K_m + \mathcal{W},$$

in which the subspace $A\mathcal{W}$ is at a gap of ϵ from a certain invariant subspace U; i.e., there exists an invariant subspace U such that

$$\Theta(U, A\mathcal{W}) = \epsilon.$$

Let P_U be any projector onto U. Then the residual $\tilde{r}$ obtained from the min-res projection process onto the augmented Krylov subspace K satisfies the inequality

$$\|\tilde{r}\|_2 \leq \min_{q \in \mathbb{P}^*_m} \left\{ \|q(A)(I - P_U)r_0\|_2 + \epsilon \, \|q(A)P_U r_0\|_2 \right\}.$$

Proof. By definition, we have

$$\|\tilde{r}\|_2 = \min_{z \in K_m + \mathcal{W}} \|r_0 - Az\|_2 \tag{3.3}$$

$$= \min_{v \in K_m, w \in \mathcal{W}} \|(r_0 - Av) - Aw\|_2. \tag{3.4}$$

Each vector v in K_m is of the form $v = s(A)r_0$, where s is a polynomial of degree $\leq m - 1$. Consequently, the vector $r_0 - Av$ is of the form $q(A)r_0$, where q belongs to the space of polynomials in $\mathbb{P}$ which satisfy the constraint $q(0) = 1$. Hence,

$$\|\tilde{r}\|_2 = \min_{q \in \mathbb{P}^*_m, w \in \mathcal{W}} \|q(A)r_0 - Aw\|_2$$

$$= \min_{q \in \mathbb{P}^*_m, w \in \mathcal{W}} \|q(A)(I - P_U)r_0 + q(A)P_U r_0 - Aw\|_2 \tag{3.5}$$

$$\leq \min_{q \in \mathbb{P}^*_m, w \in \mathcal{W}} \|q(A)(I - P_U)r_0\|_2 + \|q(A)P_U r_0 - Aw\|_2. \tag{3.6}$$

Observing that $q(A)P_U r_0$ belongs to the subspace U, the second term on the right-hand side of (3.6) is bounded from above by $\epsilon \|q(A)P_U r_0\|_2$, and this completes the proof. ∎

The above theorem can be exploited in many different ways. In particular, we may obtain different bounds depending on which type of projector P_U is used. For example, assume that P_U is the spectral projector associated with a set of eigenvalues $\lambda_1, \ldots, \lambda_s$ with $s \leq p$. Let q^*_m be the optimal GMRES polynomial obtained for the deflated initial residual $r_d = (I - P_U)r_0$:

$$\|q^*_m(A)r_d\|_2 = \min_{q \in \mathbb{P}^*_m} \|q(A)r_d\|_2.$$

Denote by $\tilde{r}_d = q^*_m(A)r_d$ the GMRES residual vector achieved on this linear system. Then, applying the theorem, we immediately get

$$\begin{aligned} \|\tilde{r}\|_2 &\leq \|q^*_m(A)r_d\|_2 + \epsilon \, \|q^*_m(A)P_U r_0\|_2 \\ &= \|\tilde{r}_d\|_2 + \epsilon \, \|q^*_m(A)P_U r_0\|_2. \end{aligned}$$

The first term in the right-hand side is the result of m steps of a GMRES iteration used to solve the deflated linear system

$$Ax = (I - P_U)r_0$$

starting with a zero initial guess. If A is diagonalizable and the initial residual has the expansion $\sum \alpha_i u_i$, the second term $q_m^*(A)P_U r_0$ will have components $q_m^*(\lambda_i)u_i\alpha_i$ in the eigenbasis. For those eigenvalues close to zero, $q_m^*(\lambda_i)$ should be close to one since $q_m^*(0) = 1$. If U is associated with eigenvalues close to zero and ϵ is small we can expect the method to behave essentially like a deflated GMRES procedure, i.e., a procedure in which the initial residual is stripped of all the components associated with the subspace U. In fact, if $\mathcal{W}$ is exactly invariant then $\epsilon = 0$ and $\|\tilde{r}\|_2 \leq \|r_d\|_2$, so we should expect the method to behave like a deflated GMRES procedure in this case. We remark that the result of Theorem 3.2 can be slightly improved by replacing the subspace $\mathcal{W}$ in the minimum (3.5) by the whole subspace K. This can be easily seen from Equation (3.4).

An immediate corollary of the theorem is the following.

COROLLARY 3.3. *Let P_U be a projector onto the invariant subspace U and let the assumption of Theorem* 3.2 *be satisfied. Also, assume that there is a polynomial q in $\mathbb{P}_m^*$ such that*

$$\|q(A)(I - P_U)r_0\|_2 \leq s_m\|(I - P_U)r_0\|_2, \tag{3.7}$$

$$\|q(A)P_U r_0\|_2 \leq c_m\|P_U r_0\|_2. \tag{3.8}$$

Then the residual $\tilde{r}$ obtained from the min-res projection process onto the augmented Krylov subspace K satisfies the inequality

$$\|\tilde{r}\|_2 \leq s_m\|(I - P_U)r_0\|_2 + \epsilon c_m\|P_U r_0\|_2, \tag{3.9}$$

and in the case when P_U is an orthogonal projector,

$$\|\tilde{r}\|_2 \leq \sqrt{s_m^2 + \epsilon^2 c_m^2}\ \|r_0\|_2. \tag{3.10}$$

The second part of the corollary follows by applying the Cauchy–Schwarz inequality to (3.9).

At this point we might provide error bounds using an eigenvector expansion of the initial residual and exploiting standard approximation theory results based on Chebyshev polynomials. These would give upper bounds for s_m and c_m from some knowledge of the spectrum of the matrix. However, these bounds would utilize in one way or another the condition number of the matrix of eigenvectors, which can be very large in case A is highly nonnormal. Therefore, this is considered only for the Hermitian case, which will be seen shortly. For the non-Hermitian case, we will consider the problem from a different angle and attempt to compare the result of the process with that of a GMRES iteration, which is expected to converge faster. This is taken up in the next section.

3.2. Comparison results. A desirable result would be that the augmented Krylov subspace method converges similarly to the GMRES algorithm applied to the deflated linear system $A\delta = r_d$. Here, the deflated residual r_d is obtained from the residual vector r_0 by removing all components in the subspace $\mathcal{W}$. In the case when $\mathcal{W}$ is an exact invariant space this turns out to be true, as was indicated above. If it is only close to an invariant subspace then an intermediate result is to be expected.

COROLLARY 3.4. *Let $\bar{r}$ be the residual obtained from m steps of GMRES applied to the $2n \times 2n$ linear system*

$$\begin{pmatrix} A & O \\ O & A \end{pmatrix} \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} = \begin{pmatrix} \epsilon P_U r_0 \\ (I - P_U)r_0 \end{pmatrix} \tag{3.11}$$

starting with a zero initial guess. Then the residual $\tilde{r}$ obtained from the min-res projection process onto the augmented Krylov subspace K satisfies the inequality

$$\|\tilde{r}\|_2 \leq \sqrt{2}\ \|\bar{r}\|_2.$$

Proof. Denote by B and $\bar{r}_0$ the matrix and right-hand side of the linear system (3.11). As is well known, the GMRES algorithm applied to system (3.11) with a zero initial guess minimizes the 2-norm $\|q(B)\bar{r}_0\|_2$ over all polynomials in $\mathbb{P}^*_m$. Let $\bar{q}$ be the polynomial which achieves this minimum. We then have

$$\begin{aligned}\|\bar{r}\|_2 &= \|\bar{q}(B)\bar{r}_0\|_2 \\ &= \left(\|\bar{q}(A)(I-P_U)r_0\|_2^2 + \|\bar{q}(A)(\epsilon P_U r_0)\|_2^2\right)^{1/2} \\ &= \left(\|\bar{q}(A)(I-P_U)r_0\|_2^2 + \epsilon^2\|\bar{q}(A)P_U r_0\|_2^2\right)^{1/2}. \qquad (3.12)\end{aligned}$$

From Theorem 3.2 we can state that

$$\|\tilde{r}\|_2 \leq\ \|\bar{q}(A)(I-P_U)r_0\|_2 + \epsilon\ \|\bar{q}(A)P_U r_0\|_2,$$

which gives the result in view of (3.12) and the inequality $|a|+|b| \leq \sqrt{2}\ \sqrt{a^2+b^2}$. ∎

In the above result we had to use a linear system of size twice that of the original matrix in order to obtain an inequality using any projector P_U. It is possible to obtain a similar comparison result using a related linear system of size n only by being more specific about the projector P_U. However, in this case, the inequality is weakened by the presence of the angle between the invariant subspace U and its complement. The following lemma will be needed.

LEMMA 3.5. *Let U and V be any two subspaces and let θ be the acute angle between them as defined by*

$$\cos\theta = \max_{u\in U,\ v\in V} \frac{|(u,v)|}{\|u\|_2\|v\|_2}.$$

Then the following inequality holds for any pair of vectors u, v with u in U and v in V:

$$\|u+v\|_2 \geq \sqrt{2}\sin\frac{\theta}{2}\left(\|u\|_2^2+\|v\|_2^2\right)^{1/2}. \qquad (3.13)$$

The proof of the lemma is straightforward and thus is omitted. If P_U is a spectral projector then it commutes with A and with any polynomial of A. In addition, $I-P_U$ is also a spectral projector which commutes with A as well as with any polynomial $q(A)$. We now show a result similar to that of Corollary 3.4.

COROLLARY 3.6. *Let P_U be the spectral projector associated with the invariant subspace U and θ the acute angle between $P_U\mathbb{C}^n$ and $(I-P_U)\mathbb{C}^n$. Let $\bar{r}$ be the residual obtained from m steps of GMRES applied to the linear system*

$$A\delta = \epsilon P_U r_0 + (I-P_U)r_0 \qquad (3.14)$$

starting with a zero initial guess. Then the residual $\tilde{r}$ obtained from the min-res projection process onto the augmented Krylov subspace K satisfies the inequality

$$\|\tilde{r}\|_2 \leq \frac{\|\bar{r}\|_2}{\sin\frac{\theta}{2}}.$$

Proof. The GMRES algorithm applied to system (3.14) with a zero initial guess minimizes the 2-norm $\|q(A)\left(\epsilon P_U r_0 + (I - P_U)r_0\right)\|_2$ over all polynomials q in $\mathbb{P}^*_m$. Let $\bar{q}$ be the polynomial which achieves this minimum. Since $\bar{q}(A)P_U r_0$ belongs to $P_U\mathbb{C}^n$ and $\bar{q}(A)(I - P_U)r_0$ belongs to $(I - P_U)\mathbb{C}^n$ we have by the previous lemma that

$$\begin{aligned}\|\bar{r}\|_2 &= \|\bar{q}(A)(I - P_U)r_0 + \bar{q}(A)(\epsilon P_U r_0)\| \\ &\geq \sqrt{2}\sin\frac{\theta}{2}\left(\|\bar{q}(A)(I - P_U)r_0\|_2^2 + \epsilon^2\|\bar{q}(A)P_U r_0\|_2^2\right)^{1/2}. \end{aligned} \tag{3.15}$$

Theorem 3.2 implies that

$$\|\tilde{r}\|_2 \leq \|\bar{q}(A)(I - P_U)r_0\|_2 + \epsilon\ \|\bar{q}(A)P_U r_0\|_2,$$

which gives the result in view of (3.15) and the inequality $|a| + |b| \leq \sqrt{2}\ \sqrt{a^2 + b^2}$. □

The angle θ is related to the conditioning of the invariant subspace U. In the ideal case when $\theta = \pi/2$, we obtain the same result as that of Corollary 3.4; namely, $\|\tilde{r}\|_2 \leq \sqrt{2}\ \|\bar{r}\|_2$.

3.3. Hermitian case. The results of the previous sections can be made more explicit in the particular case when the matrix is symmetric positive definite.

COROLLARY 3.7. *Assume that A is symmetric positive definite with eigenvalues*

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n,$$

and let the assumptions of Theorem 3.2 *be satisfied, with U being the s-dimensional eigenspace associated with the eigenvalues $\lambda_1, \ldots, \lambda_s$, where $s \leq p$. Then the residual $\tilde{r}$ obtained from the min-res projection process onto the augmented Krylov subspace K satisfies the inequality*

$$\|\tilde{r}\|_2 \leq \|r_0\|_2\ \sqrt{\frac{1}{T_m^2(\gamma)} + \epsilon^2}, \tag{3.16}$$

in which

$$\gamma \equiv \frac{\lambda_n + \lambda_{s+1}}{\lambda_n - \lambda_{s+1}}$$

and T_m is the Chebyshev polynomial of degree m of the first kind.

Proof. Define

$$\alpha = \frac{2}{\lambda_n - \lambda_{s+1}}, \quad q_m(t) = \frac{T_m(\gamma - \alpha t)}{T_m(\gamma)}.$$

Referring to the result of Corollary 3.3, we will obtain upper bounds for the numbers s_m and c_m in the corollary for the above polynomial q. Assuming that the residual r_0 is expanded in the (orthonormal) eigenbasis as

$$r_0 = \sum_{i=1}^{n} \alpha_i u_i,$$

we have

$$\|q(A)(I - P_U)r_0\|_2^2 = \frac{1}{T_m(\gamma)^2}\ \sum_{i>s} T_m(\gamma - \alpha\lambda_i)^2\alpha_i^2.$$

By the definition of α we have $|\gamma - \alpha\lambda_i| \leq 1$ for $i > s$, and as a result $|T_m(\gamma - \alpha\lambda_i)| \leq 1$. Thus, the above expression is upper bounded by

$$\|q(A)(I - P_U)r_0\|_2^2 \leq \frac{1}{T_m(\gamma)^2} \sum_{i>s} \alpha_i^2 = \frac{\|(I - P_U)r_0\|_2^2}{T_m(\gamma)^2},$$

and so we can define $s_m \equiv 1/T_m(\gamma)$. Similarly, the term $\|q(A)P_U r_0\|_2$ of Corollary 3.3 can be expanded as

$$\|q(A)P_U r_0\|_2^2 = \sum_{i \leq s} \left(q(\lambda_i)\alpha_i\right)^2.$$

In the interval $[0, \lambda_{s+1}]$ the function $q(\lambda)$ is a decreasing function and is therefore upper bounded by $q(0) = 1$. This yields

$$\|q(A)P_U r_0\|_2^2 \leq \sum_{i \leq s} \alpha_i^2 = \|P_U r_0\|_2^2.$$

As a result we can define $c_m = 1$. The result follows immediately from Corollary 3.3. □

4. Case of block-Krylov methods. Results of a slightly different type can be derived for block-Krylov methods. In these methods the subspace of projection is

$$K = K_m^{(1)} + \mathcal{W}$$

with

$$\mathcal{W} = K_m^{(2)} + K_m^{(3)} + \cdots + K_m^{(s)},$$

where $K_m^{(i)} = \text{span}[v_1^{(i)}, Av_1^{(i)}, \ldots, A^{m-1}v_1^{(i)}]$. The starting vector $v_1^{(1)}$ of the first Krylov subspace is the normalized residual $r_0/\|r_0\|_2$. A number of results for analyzing block methods have already been established in the literature [10, 14]. The approach presented here shows similar results which are somewhat simpler by introducing systematically a subsidiary approximate solution obtained by a projection step onto the subspace spanned by the initial block. Results using Chebyshev polynomials are again omitted, except in the Hermitian case.

4.1. General results. An important factor in the convergence of block methods is the subspace S spanned by the initial block, i.e., the subspace

$$S = \text{span}\{v_1^{(1)}, v_1^{(2)}, \ldots, v_1^{(s)}\}.$$

Consider any subspace U of dimension s. Typically, U will be an invariant subspace associated with the s lowest eigenvalues, but this is not required in the analysis which follows. As background, recall that any projector can be defined with the given of two subspaces, its range M, and its null space N. It is common to define N via its orthogonal complement L, which has the same dimension s as M. Thus,

$$\text{Range}(P) = M, \quad \text{Null}(P) = L^{\perp}.$$

With P is associated the decomposition of $\mathbb{C}^n$ into the direct sum

$$\mathbb{C}^n = M \oplus L^{\perp}. \tag{4.1}$$

We say that P is a projector *onto* M and *orthogonal* to L. Given two subspaces M and L, each of dimension s, a projector onto M and orthogonal to L can be defined whenever

$$M \cap L^{\perp} = \{0\},$$

which is the condition under which $\mathbb{C}^n$ is the direct sum of the two subspaces M and $L^{\perp}$. Recall also that the projection u of an arbitrary vector x onto M and orthogonal to L is defined by the requirements

$$u \in M, \quad x - u \perp L.$$

The first requirement defines the s degrees of freedom and the second defines the s constraints that allow us to extract $u = Px$ given these degrees of freedom. We now establish the following lemma.

LEMMA 4.1. *Let P_U be a projector onto a subspace U and orthogonal to a subspace L, and assume that the subspace S satisfies the condition*

$$AS \cap L^{\perp} = \{0\}. \tag{4.2}$$

Then for any vector r in $\mathbb{C}^n$ there exists a unique vector w in S such that

$$P_U(r - Aw) = 0. \tag{4.3}$$

The vector Aw is the projection of r onto the subspace AS and orthogonal to L. The vector w is the result of a projection process onto S orthogonally to L for solving the linear system $A\delta = r$ starting with a zero initial guess.

Proof. Under condition (4.2) the projector P_{AS} onto AS and orthogonal to L exists, and therefore, for any r there exists a unique Aw in AS, obtained by projecting r onto AS and orthogonally to L. This Aw satisfies the condition $r - Aw \perp L$, which implies that the vector $r - Aw$ belongs to $\text{Null}(P_U) = L^{\perp}$ or, equivalently, $P_U(r - Aw) = 0$. The rest of the proof follows from the definitions of projectors and projection methods for linear systems. ∎

Condition (4.3) can be rewritten as

$$Aw = P_U r + (I - P_U)Aw \tag{4.4}$$

because $Aw = P_U Aw + (I - P_U)Aw$ and (4.3) implies that $P_U Aw = P_U r$. The above equation means that the vector Aw has the same U-component as r in the direct sum decomposition (4.1) associated with the projector P_U. Consider the basis

$$V_1 = [v_1^{(1)}, v_1^{(2)}, \ldots, v_1^{(s)}]$$

of S. If A is nonsingular then AV_1 is a basis of AS. Let $Z = [z_1, \ldots, z_s]$ be a basis of L. Then it can easily be seen that condition (4.2) is equivalent to the nonsingularity of the $s \times s$ matrix $Z^H AS$. Condition (4.3) immediately yields

$$w = V_1(Z^H AV_1)^{-1} Z^H r.$$

THEOREM 4.2. *Let P_U be a projector onto a subspace U of dimension s such that condition* (4.2) *is satisfied for $L = \text{Null}(P_U)^{\perp}$. Let w_0 be the vector w defined by Lemma* 4.1 *for the case when $r \equiv r_0$ and denote by $\hat{r}_0$ the associated residual*

$\hat{r}_0 = r_0 - Aw_0$. *Then the residual* $\tilde{r}$ *obtained from the min-res projection process onto the augmented Krylov subspace* K *satisfies the inequality*

$$\|\tilde{r}\|_2 \leq \min_{p \in \mathbb{P}^*_m} \|q(A)(I - P_U)\ \hat{r}_0\|_2. \tag{4.5}$$

Proof. We start similarly to the proof of Theorem 3.2:

$$\begin{aligned}\|\tilde{r}\|_2 &= \min_{z \in K = K_m + \mathcal{W}} \|r_0 - Az\|_2 \\ &= \min_{v \in K_m,\ w \in K} \|(r_0 - Av) - Aw\|_2.\end{aligned}$$

As was seen before, a generic vector $r_0 - Av$ is of the form $q(A)r_0$, where q is a polynomial of degree $\leq m$ such that $q(0) = 1$, and therefore

$$\begin{aligned}\|\tilde{r}\|_2 &= \min_{q \in \mathbb{P}^*_m,\ w \in K} \|q(A)r_0 - Aw\|_2 \\ &= \min_{q \in \mathbb{P}^*_m,\ w \in K} \|q(A)(I - P_U)r_0 + q(A)P_U r_0 - Aw\|_2.\end{aligned}$$

For any polynomial q in $\mathbb{P}^*_m$ and for any vector w in K we have

$$\|\tilde{r}\|_2 \leq \|q(A)(I - P_U)r_0 + q(A)P_U r_0 - Aw\|_2. \tag{4.6}$$

Now consider the particular vector $w = q(A)w_0$, where the vector w_0 is defined by the theorem. Using the result of Lemma 4.1 and equality (4.4) we obtain

$$\begin{aligned}q(A)P_U r_0 - Aw &= q(A)P_U r_0 - Aq(A)w_0 \\ &= q(A)P_U r_0 - q(A)Aw_0 \\ &= q(A)P_U r_0 - q(A)[P_U r_0 + (I - P_U)Aw_0] \\ &= -q(A)(I - P_U)Aw_0.\end{aligned}$$

Substituting this in Equation (4.6) for any polynomial q results in

$$\|\tilde{r}\|_2 \leq \|q(A)(I - P_U)(r_0 - Aw_0)\|_2. \tag{4.7}$$

Taking the minimum of the right-hand side over all polynomials in $\mathbb{P}^*_m$ yields the desired result. ∎

This simple theorem states that a block-GMRES method will do at least as well as a GMRES method on the linear system whose initial residual has been stripped of the components in the subspace U by a projection process on the initial subspace S. The removal of these undesired components is achieved by a projection process onto S orthogonally to $L = \text{Null}(P_U)^\perp$, as expressed by the Galerkin conditions

$$w_0 \in S, \qquad r_0 - Aw_0 \perp \text{Null}(P_U)^\perp.$$

Note again that P_U is any projector onto the subspace U.

The projector $I - P_U$ in Equation (4.5) is not really needed since $\hat{r}_0$ has no components in the subspace U, and so $(I - P_U)\hat{r}_0 = \hat{r}_0$. However, its presence is helpful when P_U is a spectral projector, since in this situation

$$q(A)(I - P_U) = q((I - P_U)A(I - P_U)),$$

showing that the GMRES iteration associated with the minimum in (4.5) is equivalent to a GMRES iteration for solving a linear system *restricted to the spectral complement of* U.

4.2. Block-Krylov methods in the symmetric positive definite (SPD) case. We assume throughout this section that A is SPD with the eigenvalues

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n.$$

Here, the subspace U is chosen to be the invariant subspace associated with the eigenvalues $\lambda_1, \ldots, \lambda_p$ and P_U is the spectral projector associated with U. In this case, P_U is the *orthogonal* projector onto U and the subspace L, which was defined as the orthogonal complement of the null space of P, becomes equal to U itself.

By selecting the polynomial in Theorem 4.2 carefully a rather simple result can be obtained.

THEOREM 4.3. *Let P_U be the orthogonal projector onto the invariant subspace associated with the eigenvalues $\lambda_1, \ldots, \lambda_p$ and assume condition* (4.2) *is satisfied. Let w_0 be the vector w defined by Lemma* 4.1 *for the case when $r \equiv r_0$ and $\hat{r}_0 = r_0 - Aw_0$. Then the residual $\tilde{r}$ obtained from the min-res projection process onto the augmented Krylov subspace K satisfies the inequality*

$$\|\tilde{r}\|_2 \leq \frac{\|\hat{r}_0\|_2}{T_m(\gamma)} \tag{4.8}$$

with

$$\gamma \equiv \frac{\lambda_n + \lambda_{p+1}}{\lambda_n - \lambda_{p+1}}.$$

Proof. According to Theorem 4.2, for any polynomial q in $\mathbb{P}^*_m$ we have

$$\|\tilde{r}\|_2 \leq \|q(A)(I - P_U)\ \hat{r}_0\|_2 \leq \|q(A)(I - P_U)\|_2\ \|\hat{r}_0\|_2. \tag{4.9}$$

Since $I - P_U$ is a spectral projector of A we have

$$q(A)(I - P_U) = (I - P_U)q(A) = (I - P_U)q(A)(I - P_U).$$

The only nonzero eigenvalues of the Hermitian operator $(I - P_U)q(A)(I - P_U)$ are $q(\lambda_i)$ for $i > p$. Thus,

$$\|(I - P_U)q(A)(I - P_U)\|_2 = \max_{i=p+1,\ldots,n} |q(\lambda_i)|. \tag{4.10}$$

Consider the polynomial $q_m(t)$ defined by

$$q_m(t) = \frac{T_m(\gamma - \alpha t)}{T_m(\gamma)},$$

where γ is defined above and

$$\alpha \equiv \frac{2}{\lambda_n - \lambda_{p+1}}.$$

Clearly, q_m belongs to $\mathbb{P}^*_m$. In addition, for t in the closed interval $[\lambda_{p+1}, \lambda_n]$ we have $|\gamma - \alpha t| \leq 1$ so that $|T_m(\gamma - \alpha t)| \leq 1$. For this polynomial the norm of the Hermitian operator $(I - P_U)q(A)(I - P_U)$ in (4.10) becomes

$$\|q(A)(I - P_U)\|_2 = \|(I - P_U)q(A)(I - P_U)\|_2 = \max_{i=p+1,\ldots,n} |q_m(\lambda_i)| \leq \frac{1}{T_m(\gamma)}. \tag{4.11}$$

Substituting this inequality in (4.9) yields the desired result. ∎

5. Numerical experiment. The behaviors of the deflated algorithms and the block-GMRES algorithms are now illustrated by a simple example. Consider a diagonal matrix of size $n = 200$ whose diagonal entries are given by

$$d_i = \begin{cases} \frac{i}{n} & \text{when } i > 4, \\ 0.05 \times \frac{i}{n} & \text{when } i \leq 4. \end{cases}$$

This distribution is chosen to have a small cluster of eigenvalues around the origin. In all tests, the right-hand side b of the linear system is made of (the same) pseudo-random values and the initial guess taken is the zero vector. Though the matrix is symmetric, nonsymmetric iterative solvers such as GMRES and block-GMRES are used in this experiment. The following runs were made.

1. Standard GMRES without restarts and restarted GMRES, with a Krylov dimension of 40.
2. Block-GMRES without restarts. The block size chosen is four, which is the size of the cluster.
3. A deflated GMRES algorithm as described in [9] and [2]. This consists of adding approximate eigenvectors obtained from the previous Arnoldi step to the Krylov subspace. The test uses a subspace dimension of 40, the last four of which are approximate eigenvectors (except in the first outer iteration). This is denoted by dGMRES$(40, 4)$.
4. For comparison, a run of (nonrestarted) GMRES is shown on the deflated system. This system of dimension 196 has a diagonal coefficient matrix with entries $d_5, d_6, \ldots, d_{200}$ and the right-hand side b with components $b_5, \ldots, b_{200}$. A zero initial guess was also used.

In the block-GMRES case, four linear systems are actually solved simultaneously, the first of which is the desired linear system. The right-hand sides of the other three linear systems are chosen randomly and the associated initial guesses are again zero vectors.

The convergence history for these runs is plotted in Figure 5.1. As observed, all curves, except the restarted GMRES curve, have similar convergence slopes toward the final phase of the iteration. The first 40 steps of GMRES, GMRES(40), and dGMRES$(40, 4)$ (deflated GMRES) are identical. Differences appear at around step 60, halfway into the second outer loop between full GMRES and the other two methods. GMRES(40) and dGMRES$(40, 4)$ are still identical until step 76. Indeed, in the first outer loop there is no eigenspace information to be fed into dGMRES, so a plain restarted GMRES is used. The last four vectors entered into dGMRES are eigenvectors obtained from the first Krylov subspace. Then the behavior of the iteration from that point on is very close to that of the full GMRES and GMRES on the deflated system.

It is interesting to note that in this case the full GMRES algorithm performs best. We must keep in mind that after step 40 the full GMRES iteration uses a subspace which includes the same eigenvectors as dGMRES$(40, 4)$. It is therefore able to capture those eigenmodes in the same way as the deflated GMRES, as shown by the curves. Also interesting is the observation that the block-GMRES algorithm seems to take longer to capture the cluster and reach the final convergence phase. If we had to solve four simultaneous linear systems, the block-GMRES algorithm would be competitive since it would take an average of 45 steps for each linear system to converge (assuming they converge at roughly equal speed on average). If we had only one linear system to solve, the results of the plot indicate that a plain or a deflated

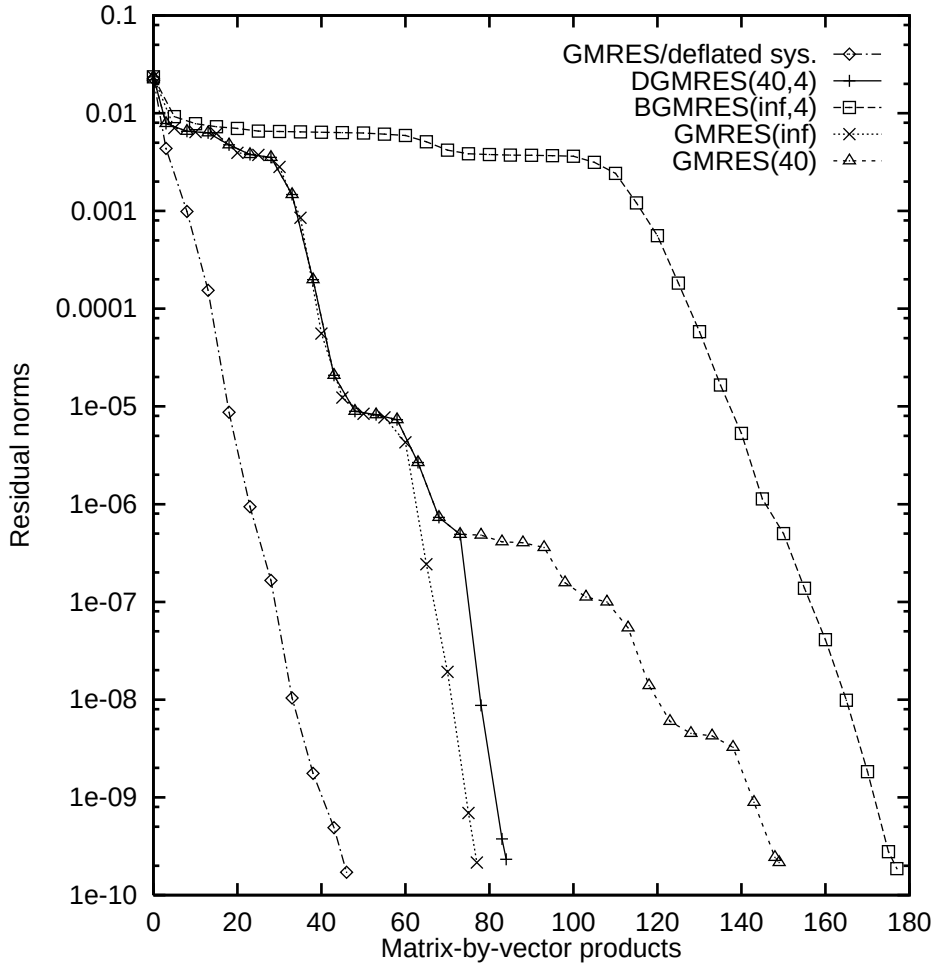


FIG. 5.1. *Behavior of GMRES and block-GMRES on a matrix whose spectrum has a cluster around the origin.*

GMRES run may achieve far better performance. This is confirmed by experiments elsewhere; see, e.g., [2].

Acknowledgment. I wish to thank the members of the Applied Mathematics group at UCLA for their hospitality during my visit in the spring of 1995. The Minnesota Supercomputer Institute provided computer facilities and other resources to conduct this research.

REFERENCES

[1] T. F. CHAN AND W. L. WAN, *Analysis of Projection Methods for Solving Linear Systems with Multiple Right-hand Sides*, Technical report CAM-94-26, University of California at Los Angeles, Department of Mathematics, Los Angeles, CA, 1994.

[2] A. CHAPMAN AND Y. SAAD, *Deflated and augmented Krylov subspace techniques*, Numer. Linear Algebra Appl., to appear.

[3] F. CHATELIN, *Spectral Approximation of Linear Operators*, Academic Press, New York, 1984.

[4] J. Erhel, K. Burrage, and B. Pohl, *Restarted GMRES Preconditioned by Deflation*, Technical report, IRISA, Rennes, France, 1994; J. Comput. Appl. Math., 1995, to appear.
[5] C. Farhat, L. Crivelli, and F. X. Roux, *Extending Substructure Based Iterative Solvers to Multiple Load and Repeated Analyses*, Technical report, Center for Space Structures and Controls, Boulder, CO, 1993.
[6] P. F. Fischer, *Projection Techniques for Iterative Solution of $Ax = b$ with Successive Right-hand Sides*, Technical report TR-93-90, ICASE, Hampton, VA, 1993.
[7] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1965.
[8] S. A. Kharchenko and A. Yu. Yeremin, *Eigenvalue translation based preconditioners for the GMRES(k) method*, Numer. Linear Algebra Appl., 2 (1995), pp. 51–70.
[9] R. B. Morgan, *A restarted GMRES method augmented with eigenvectors*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1154–1171.
[10] D. O'Leary, *The block conjugate gradient algorithm and related methods*, Linear Algebra Appl., 29 (1980), pp. 243–322.
[11] K. Guru Prasad, D. E. Keyes, and J. H. Kane, *GMRES for Sequentially Multiple Nearby Systems*, Technical report, Old Dominion University, 1995.
[12] Y. Saad, *A flexible inner-outer preconditioned GMRES algorithm*, SIAM J. Sci. Stat. Comput., 14 (1993), pp. 461–469.
[13] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS publishing, New York, 1996.
[14] V. Simoncini and E. Gallopoulos, *Convergence properties of block GMRES and matrix polynomials*, Linear Algebra Appl., to appear.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, pp. 450–463, April 1997

014

ON THE COMPLEXITY OF MATRIX BALANCING*

B. KALANTARI†, L. KHACHIYAN†, AND A. SHOKOUFANDEH†

Abstract. An $n \times n$ matrix with nonnegative entries is said to be *balanced* if for each $i = 1, \ldots, n$ the sum of the entries of its ith row is equal to the sum of the entries of its ith column. An $n \times n$ matrix A with nonnegative entries is said to be *balancable via diagonal similarity scaling* if there exists a diagonal matrix X with positive diagonal entries such that XAX^{-1} is balanced. We give upper and lower bounds on the entries of X and prove the necessary sensitivity analysis in the required accuracy of the minimization of an associated convex programming problem. These results are used to prove the polynomial-time solvability of computing X to any prescribed accuracy.

Key words. nonnegative matrices, matrix balancing, line-sum-symmetric scaling, polynomial-time solvability, matrix preconditioning, eigenvalue problems

AMS subject classifications. 65F35, 90C25, 68A20

PII. S0895479895289765

1. Introduction.

An $n \times n$ matrix A with nonnegative entries is said to be *balanced* if for each $i = 1, \ldots, n$ the sum of the entries of its ith row is equal to the sum of the entries of its ith column:

$$A\mathbf{1} = A^T\mathbf{1}, \tag{1.1}$$

where $\mathbf{1}$ is the n-vector of all ones. A is said to be *balancable via diagonal similarity scaling* (or simply balancable) if there exists a diagonal matrix X with positive diagonal entries such that XAX^{-1} is balanced; i.e.,

$$XAX^{-1}\mathbf{1} = X^{-1}A^TX\mathbf{1}. \tag{1.2}$$

The matrix balancing problem can be defined more generally: An $n \times n$ matrix $B = (b_{ij})$ with arbitrary real entries is said to be *balanced in the l_p-norm* ($p > 0$) if for each $i = 1, \ldots, n$ its ith row and column have the same l_p-norm. An invertible diagonal matrix $Y = \mathrm{diag}(y_1, \ldots, y_n)$ *balances B in the l_p-norm* if for each $i = 1, \ldots, n$ the l_p-norm of the ith row and column of YBY^{-1} are identical; i.e.,

$$\sum_{j=1}^{n} \left| b_{ij}\frac{y_i}{y_j} \right|^p = \sum_{j=1}^{n} \left| b_{ji}\frac{y_j}{y_i} \right|^p, \quad i = 1, \ldots, n. \tag{1.3}$$

Clearly, an invertible diagonal matrix $Y = \mathrm{diag}(y_1, \ldots, y_n)$ balances B in the l_p-norm if and only if the positive diagonal matrix $X = \mathrm{diag}(|y_1|^p, \ldots, |y_n|^p)$ balances the nonnegative matrix $A = (|b_{ij}|^p)$ in the l_1-norm. The general matrix balancing problem in the l_p-norm can thus be reduced to the case of nonnegative matrix balancing via a positive diagonal matrix (see, e.g., [8]).

Osborne [7] considered the case of $p = 2$ and its application in preconditioning a given matrix B in order to increase the accuracy of the computation of its eigenvalues

* Received by the editors August 1, 1995; accepted for publication (in revised form) by M. L. Overton June 11, 1996. This research was supported in part by National Science Foundation grant CCR-9208371.

http://www.siam.org/journals/simax/18-2/28976.html

† Department of Computer Science, Rutgers University, New Brunswick, NJ 08903 (kalantar@cs.rutgers.edu).

(B and YBY^{-1} have the same set of eigenvalues). Through an iterative process, Osborne showed that if $Y^* = \mathrm{diag}(y_1^*, \ldots, y_n^*)$ satisfies (1.3), then the vector $y^* = (y_1^*, \ldots, y_n^*)$ is the minimizer of the function

$$\phi_B(y) = \left(\sum_{i,j=1}^{n} \left| b_{ij} \frac{y_i}{y_j} \right|^2 \right)^{1/2},$$

where the minimization ranges over all invertible $Y = \mathrm{diag}(y_1, \ldots, y_n)$. Conversely, the minimizer of $\phi_B(y)$, if it exists, satisfies (1.3). Note that $\phi_B(y)$ is the Frobenius norm of the matrix YBY^{-1} satisfying

$$\frac{1}{\sqrt{n}} \phi_B(y) \leq \|YBY^{-1}\| \leq \phi_B(y).$$

Henceforth, $\|\cdot\|$ denotes the l_2-norm. In view of the above inequality and Osborne's result, the balancing of the nonnegative matrix $A = (|b_{ij}|^2)$ also bounds the quantity

$$\nu(B) = \inf\{\|YBY^{-1}\| : y \in \Re^n, \quad Y = \mathrm{diag}(y_1, \ldots, y_n), \quad \Pi_{i=1}^n y_i \neq 0\}.$$

For a description of $\nu(B)$ as a generalized eigenvalue problem see Boyd et al. [1].

The problem of nonnegative matrix balancing has been examined by several researchers. Balancability has been called *line-sum-symmetric scaling* (see Eaves et al. [2]) and *balancing* (see Grad [3], Rothblum and Schneider [8], and Schneider and Zenios [9]). Characterization theorems on nonnegative balancable matrices have been given by Osborne [7] and Eaves et al. [2]. Other results on matrix balancing, including applications and iterative algorithms, have been given by Osborne [7], Grad [3], and Schneider and Zenios [9].

From now on we shall consider nonnegative matrices, and we shall say a nonnegative matrix is *balanced* to mean that it is balanced in the sense of (1.1) and (1.2). In this paper we prove the polynomial-time solvability of the problem of balancing a nonnegative matrix to any prescribed accuracy.

Clearly, without loss of generality we may assume that a given $n \times n$ nonnegative matrix $A = (a_{ij})$ satisfies $a_{ii} = 0$ for all $i = 1, \ldots, n$. Corresponding to such a matrix A there exists a directed graph $G_A = (V, E)$, where $V = \{1, \ldots, n\}$ and where $E = \{(i, j) : a_{ij} > 0\}$. Without loss of generality we may also assume that G_A, when viewed as an undirected graph, is connected. Otherwise, after a permutation of $V = \{1, \ldots, n\}$ the given matrix A can be replaced by $\mathrm{diag}(A_1, \ldots, A_r)$, where each of $A_1, \ldots, A_r$ is a square matrix whose corresponding directed graph is connected. Thus A is balancable if and only if $A_1, \ldots, A_r$ are balancable. Moreover, it can be shown that for each $i = 1, \ldots, r$ A_i is balancable if and only if the corresponding graph G_{A_i} is strongly connected (Theorem 1).

DEFINITION 1. *Given a positive tolerance ε, a nonnegative $n \times n$ matrix A is said to be* balanced to the absolute error of ε *if* $\|A\mathbf{1} - A^T\mathbf{1}\| \leq \varepsilon$. *A positive diagonal matrix X is said to* balance A to the absolute error *of ε if the matrix XAX^{-1} is balanced to the absolute error of ε:*

$$\|XAX^{-1}\mathbf{1} - X^{-1}A^TX\mathbf{1}\| \leq \varepsilon. \tag{1.4}$$

DEFINITION 2. *Given a positive tolerance ε, a nonnegative $n \times n$ matrix A is said to be* balanced to the relative error of ε *if* $\|A\mathbf{1} - A^T\mathbf{1}\|/\mathbf{1}^TA\mathbf{1} \leq \varepsilon$. *A positive*

diagonal matrix X is said to balance A to the relative error *of ε if the matrix XAX^{-1} is balanced to the relative error of ε:*

$$\frac{\|XAX^{-1}\mathbf{1} - X^{-1}A^TX\mathbf{1}\|}{\mathbf{1}^TXAX^{-1}\mathbf{1}} \leq \varepsilon. \tag{1.5}$$

Let

$$f(x) = \sum_{i,j=1}^{n} a_{ij}\frac{x_i}{x_j}.$$

Then any stationary point $x > 0$ of f yields an exact balancing, $X = \mathrm{diag}(x)$. Since the optimization of $f(x)$ is a standard geometric programming problem, applying the change of variable

$$x = (e^{w_1}, \ldots, e^{w_n})^T \in \Re^n,$$

we obtain the convex function

$$F(w) = \sum_{i,j=1}^{n} a_{ij}e^{w_i - w_j},$$

each exact minimizer $w^* = (w_1^*, \ldots, w_n^*) \in \Re^n$ of which yields an exact balancing $X(w^*) = \mathrm{diag}(e^{w_1^*}, \ldots, e^{w_n^*})$ of A. It is easy to see that if $\|\nabla F(w)\| \leq \varepsilon$, then $X(w)$ satisfies (1.4), and vice versa. Moreover, letting $G(w) = \ln F(w)$, we see that $\|\nabla G(w)\| \leq \varepsilon$ if and only if $X(w)$ satisfies (1.5).

In this paper we show the polynomial-time solvability of the balancing problem. Specifically, we prove the following complexity result (Theorem 5):

Let A be an $n \times n$ nonnegative matrix, $a_{ii} = 0$, for all $i = 1, \ldots, n$. Suppose that $G_A = (V, E)$ is strongly connected. Let $a_{\min} = \min\{a_{ij} : (i,j) \in E\}$, $\sigma = \sum_{(i,j)\in E} a_{ij}$, and $v = a_{\min}/\sigma$. For any given accuracy $\varepsilon \in (0,1)$, in $O(n^4 \ln(\frac{n}{\varepsilon}\ln\frac{1}{v}))$ arithmetic operations over $O(\ln(\frac{n}{\varepsilon v}))$-bit numbers, we can compute a positive diagonal matrix $X = \mathrm{diag}(e^{w_1}, \ldots, e^{w_n})$ so that XAX^{-1} is balanced to the relative error of ε and the absolute error of $e\sigma\varepsilon$.

In order to obtain the above result we first state a characterization on balancable matrices, Theorem 1. In particular, this theorem implies that an arbitrary nonnegative matrix is balancable if and only if its corresponding graph is the union of strongly connected graphs.

In Theorem 2 we prove the following:

An $n \times n$ nonnegative matrix A with $a_{ii} = 0$, for all $i = 1, \ldots, n$, and G_A strongly connected can be balanced by a diagonal matrix $X^* = \mathrm{diag}(e^{w_1^*}, \ldots, e^{w_n^*})$ such that

$$|w_i^*| \leq \frac{n-1}{2}\ln\frac{e}{(n-1)v}, \quad i = 1, \ldots, n.$$

We give an example of ill-behaved balancable matrices for which the above bound on balancing factors is sharp up to a constant factor.

To obtain the necessary bound on the accuracy of the minimization of $G(w)$ we prove the following in Theorem 3 and Corollary 2:

For any given $\varepsilon \in (0,1)$, the minimization of $G(w)$ to an absolute accuracy of $\delta = \varepsilon^2/16$ gives a point $(w_1, \ldots, w_n)$ so that (1.5) is satisfied with $X(w) =$

$\mathrm{diag}(e^{w_1}, \ldots, e^{w_n})$; *i.e., if* w^* *is an exact minimizer of* $G(w)$ *in* $\Re^n$, $G(w) - G(w^*) \leq \delta$ *implies that* $X(w)$ *balances* A *to a relative error of* ε.

Again, we give a simple example for which the above bound is optimal up to a constant factor. In Theorem 4 we prove the following:

For any given $\varepsilon \in (0,1)$, $\|w - w^*\| \leq \varepsilon^2/(16\sqrt{2})$ *implies* $G(w) - G(w^*) \leq \delta = \varepsilon^2/16$.

The above results will imply the polynomial-time solvability of the problem of balancing to any prescribed relative or absolute error via the ellipsoid algorithm or interior-point Newton methods; see, e.g., Nesterov and Nemirovskii [6].

The remainder of the paper is organized as follows. In section 2 we state a characterization result on balancable matrices. In section 3 we derive our bounds on balancing matrices. In section 4 we bound the required absolute accuracy of the minimization of $G(w)$. Finally, in section 5 we prove the polynomial-time solvability of the balancing problem.

2. Characterization of nonnegative balancable matrices. The following characterization of balancable matrices is due to Eaves et al. [2]. The equivalence of conditions (ii) and (iii) is essentially due to Osborne [7]. For the sake of completeness we provide a proof of this theorem. The proof can be viewed as an alternative proof to that of [2].

THEOREM 1. *Let* A *be an* $n \times n$ *nonnegative matrix,* $a_{ii} = 0$, *for all* $i = 1, \ldots, n$, *whose graph* $G_A = (V, E)$ *when viewed as an undirected graph is connected. The following statements are equivalent.*

(i): G_A *is strongly connected.*

(ii): $f(x) = \sum_{(i,j)\in E} a_{ij} \frac{x_i}{x_j}$ *attains its infimum over* $\Omega = \{x \in \Re^n : \sum_{i=1}^n x_i = 1,\ x_i > 0,\ \ i = 1, \ldots, n\}$.

(iii): A *is balancable.*

(iv): *There exist* $n \times n$ *circuit matrices* $C_1, \ldots, C_q$ *(i.e., each* C_k *is a matrix with* 0-1 *entries whose graph* G_k *is a simple cycle through* $n_k \leq n$ *vertices) and a positive diagonal matrix* $X^* = \mathrm{diag}(x_1^*, \ldots, x_n^*)$ *such that*

$$X^* A {X^*}^{-1} = \sum_{k=1}^{q} \alpha_k C_k$$

with some positive weights $\alpha_1, \ldots, \alpha_q$.

Proof. (i)⇒(ii): Let $\bar{x}$ be a boundary point of Ω. Then $\bar{x_i} > 0$ and $\bar{x_j} = 0$ for some $i, j \in \{1, \ldots, n\}$. Since G_A is strongly connected, there exist edges $(i_1 = i, i_2), (i_2, i_3), \ldots, (i_{k-1}, i_k = j) \in E$, where $i_1, \ldots, i_k$ are distinct. Then

$$f(x) \geq \sum_{l=1}^{k-1} a_{i_l i_{l+1}} \frac{x_{i_l}}{x_{i_{l+1}}}.$$

As $x \in \Omega$ approaches $\bar{x}$, the right-hand side of the above approaches $+\infty$. From this, the fact that Ω is bounded, and since f is continuous and positive on Ω, it follows that f attains its infimum.

(ii) ⇒ (iii): Let $x^* \in \Omega$ be a minimizer of f over Ω. Since f is homogeneous of degree zero, x^* minimizes f over $\{x : x > 0\}$. Hence $\nabla f(x^*) = 0$, which gives

$$\sum_{j=1}^{n} a_{ij} \frac{x_i^*}{x_j^*} = \sum_{j=1}^{n} a_{ji} \frac{x_j^*}{x_i^*}.$$

(iii)⇒(iv): Let $A^* = X^* A X^{*^{-1}}$. Since A^* is balanced, $A^* \in K$, where

$$K = \left\{ Y = (y_{ij}) \geq 0 : \sum_{j=1}^{n} y_{ij} - \sum_{j=1}^{n} y_{ji} = 0, \quad i = 1, \ldots, n, \quad y_{ij} = 0 \ for \ (i,j) \notin E \right\}.$$

So A^* is a positive combination of generators of the cone K. But the graph corresponding to any generator of K is a simple cycle, proving (iv).

(iv)⇒(i): From (iv), G_A can be decomposed as the union of directed cycles. Given this and the fact that G_A is connected, it is easy to argue the reachability from any vertex i to any other vertex j. □

COROLLARY 1. *Letting* $A_k = X^{*^{-1}} C_k X^*$, *we have*

$$A = \sum_{k=1}^{q} \alpha_k A_k,$$

where X^* *simultaneously balances each of the* A_k*'s.*

Remark 1. Theorem 1 implies that the balancability of A can be tested in $O(|E|)$ time.

Remark 2. It can be shown that under the assumption of Theorem 1 the balancing matrix X^* is unique up to a scalar factor (see Osborne [7] and Eaves et al. [2]).

3. Bounds on balancing matrix.

THEOREM 2. *Let* A *be an* $n \times n$ *nonnegative matrix,* $a_{ii} = 0$, *for all* $i = 1, \ldots, n$, *and* G_A *strongly connected. There exists a positive diagonal* $X^* = \mathrm{diag}(x_1^*, \ldots, x_n^*)$ *balancing* A *such that*

$$\left[\frac{e}{(n-1)v}\right]^{\frac{1-n}{2}} < x_i^* < \left[\frac{e}{(n-1)v}\right]^{\frac{n-1}{2}}, \quad i = 1, \ldots, n, \tag{3.1}$$

where

$$v = \frac{a_{\min}}{\sigma}, \quad a_{\min} = \min\{a_{ij} \ : (i,j) \in E\}, \quad \sigma = \sum_{(i,j)\in E} a_{ij}. \tag{3.2}$$

Proof. From Theorem 1, the minimizer x^* of $f(x)$ over $\{x : x > 0\}$ exists. Let $X^* = \mathrm{diag}(x^*)$. From (3.2) and the optimality of x^* we have

$$a_{\min} \sum_{(i,j)\in E} \frac{x_i^*}{x_j^*} \leq \sum_{(i,j)\in E} a_{ij} \frac{x_i^*}{x_j^*} = f(x_1^*, \ldots, x_n^*) \leq f(1, \ldots, 1) = \sigma. \tag{3.3}$$

Suppose without loss of generality that

$$\frac{\max\{x_1^*, \ldots, x_n^*\}}{\min\{x_1^*, \ldots, x_n^*\}} = \frac{x_1^*}{x_2^*}.$$

Since G_A is strongly connected, there exists a simple directed path from 1 to 2, say $P = (i_1, i_2), \ldots, (i_{t-1}, i_t)$, where $i_1 = 1$, $i_t = 2$, and $t \leq n$. From (3.3) and the arithmetic–geometric mean inequality we get

$$\begin{aligned} \frac{x_1^*}{x_2^*} = \prod_{(i,j)\in P} \frac{x_i^*}{x_j^*} &\leq \left[\frac{1}{(t-1)} \sum_{(i,j)\in P} \frac{x_i^*}{x_j^*}\right]^{t-1} \leq \left[\frac{1}{(t-1)v}\right]^{t-1} \\ &< \left[\frac{e}{(t-1)v}\right]^{t-1} \leq \left[\frac{e}{(n-1)v}\right]^{n-1}, \end{aligned}$$

where the last inequality follows from $t \le n$ and $v \le 1/n$. Replacing X^* by tX^* with $t = 1/\sqrt{x_1^* x_2^*}$, (3.1) follows. □

We now give an example of nonnegative balancable matrices which are ill behaved. Consider $n = 2k+1$ and $\varepsilon \in (0,1)$. Let A be an $n \times n$ nonnegative matrix with the following positive entries:

$$a_{i,i+1} = a_{2k+2-i,2k+1-i} = 1, \tag{3.4}$$

$$a_{i+1,i} = a_{2k+1-i,2k+2-i} = \varepsilon, \tag{3.5}$$

for $i = 1, \ldots, k$, and

$$a_{n1} = a_{1n} = 1. \tag{3.6}$$

Observe that the graph G_A of matrix A can be decomposed into two cycles through its n vertices. Next, define the diagonal matrix $X^* = \operatorname{diag}(x_1^*, \ldots, x_n^*)$ as follows:

$$x_{2k+2-i}^* = x_i^* = \varepsilon^{\frac{k+2-2i}{4}}, \quad i = 1, \ldots, k+1. \tag{3.7}$$

To illustrate (3.4)–(3.7), consider the case $k = 3$:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ \varepsilon & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & \varepsilon & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \varepsilon & 0 & \varepsilon & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & \varepsilon & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & \varepsilon \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

$$X^* = \operatorname{diag}(\varepsilon^{\frac{3}{4}}, \varepsilon^{\frac{1}{4}}, \varepsilon^{-\frac{1}{4}}, \varepsilon^{-\frac{3}{4}}, \varepsilon^{-\frac{1}{4}}, \varepsilon^{\frac{1}{4}}, \varepsilon^{\frac{3}{4}}).$$

It is easy to check that the positive entries of $X^* A X^{*^{-1}}$ are given by

$$a_{i,i-1}^* = a_{i-1,i}^* = \sqrt{\varepsilon}, \quad i = 2, \ldots, n,$$

$$a_{1n}^* = a_{n1}^* = 1.$$

From this it follows that

$$\sum_{j=1}^{n} a_{ij}^* = \sum_{j=1}^{n} a_{ji}^* = 2\sqrt{\varepsilon}, \quad i = 2, \ldots, n-1$$

and

$$\sum_{j=1}^{n} a_{ij}^* = \sum_{j=1}^{n} a_{ji}^* = 1 + \sqrt{\varepsilon}, \quad i = 1, n.$$

Thus X^* balances A and

$$\frac{\max\{x_1^*, \ldots, x_n^*\}}{\min\{x_1^*, \ldots, x_n^*\}} = \left(\frac{1}{\varepsilon}\right)^{(n-1)/4}.$$

Since X^* is unique up to scalar multiplication, the above bound holds for any balancing of A. In comparison, Theorem 1 gives

$$\frac{\max\{x_1^*,\ldots,x_n^*\}}{\min\{x_1^*,\ldots,x_n^*\}} < \left[\frac{e}{n-1}\cdot\frac{(n-1)\varepsilon+n+1}{\varepsilon}\right]^{n-1} \sim \left[\frac{e}{\varepsilon}\right]^{n-1}.$$

Remark 3. It follows from the proof of Theorem 2 that (3.1) can be strengthened as follows:

$$\left[\frac{e}{dv}\right]^{\frac{-d}{2}} < x_i^* < \left[\frac{e}{dv}\right]^{\frac{d}{2}}, \quad i=1,\ldots,n,$$

where d is the longest (directed) shortest path between a pair of vertices in G_A.

4. A sensitivity theorem for the convex program. For a given $n\times n$ nonnegative matrix A, consider the function

$$G(w) = \ln F(w),$$

where

$$F(w) = f(e^{w_1},\ldots,e^{w_n}) = \sum_{(i,j)\in E} a_{ij}e^{w_i-w_j}.$$

THEOREM 3. *Let A be an $n\times n$ nonnegative matrix, $a_{ii}=0$, for all $i=1,\ldots,n$, and G_A strongly connected. Let $w^*\in\Re^n$ be an exact minimizer of $G(w)$. Suppose that for a given $\delta>0$ we have $w\in\Re^n$ satisfying*

$$G(w)-G(w^*)\le\delta.$$

Then

$$\|\nabla G(w)\| \le 2\sqrt{e^{2\delta}-1}.$$

Before proving the theorem we state the following corollary, which gives the required accuracy of the optimization of $G(w)$. From Theorem 3 and the fact that $x/2<\ln(1+x)$ for all $x\in(0,1)$ we get the following corollary.

COROLLARY 2. *Given $\varepsilon\in(0,1)$, let*

$$\delta=\frac{\varepsilon^2}{16}.$$

If $w\in\Re^n$ satisfies $G(w)-G(w^*)\le\delta$, then $\|\nabla G(w)\|\le\varepsilon$ (and hence the diagonal matrix $X(w)=\operatorname{diag}(e^{w_1},\ldots,e^{w_n})$ balances A to the relative error of ε). □

Observe that the bound of Corollary 2 is optimal up to a constant factor: the identity matrix balances A to the relative error of ε, where

$$A=\begin{pmatrix} 0 & \frac{1}{2}+\frac{\varepsilon}{2\sqrt{2}} \\ \frac{1}{2}-\frac{\varepsilon}{2\sqrt{2}} & 0\end{pmatrix}$$

and

$$\delta = G(0)-G(w^*) = -\frac{1}{2}\ln\left(1-\frac{\varepsilon^2}{2}\right)\sim\frac{\varepsilon^2}{4}.$$

In order to prove Theorem 3 we need some auxiliary lemmas.

LEMMA 1. *For any given number $a \geq 1$ and $x \in \Re^n$ define*

$$\sigma(x,a) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - a)^2}{n(n-1)}}.$$

Let

$$B(n,a) = \max\left\{\sigma(x,a) : \frac{1}{n}\sum_{i=1}^{n} x_i = a, \quad \prod_{i=1}^{n} x_i = 1, \quad x > 0\right\}.$$

The above optimization problem has an optimal solution $(x_1, \ldots, x_n)$ such that $x_1 = \cdots = x_{n-1} \leq x_n$.

Proof. For $n = 1, 2$ there is nothing to prove. We first prove the lemma for $n = 3$. Consider

$$\max\left\{\sum_{i=1}^{3}(x_i - a)^2 : \frac{1}{3}\sum_{i=1}^{3} x_i = a, \quad \prod_{i=1}^{3} x_i = 1, \quad x_i > 0, \quad i = 1,2,3\right\}.$$

Since $a \geq 1$, from the relationship between arithmetic–geometric means the feasible region is nonempty. The optimality condition gives

$$2(x_i - a) = \lambda_1 + \frac{\lambda_2}{x_i}, \quad i = 1,2,3,$$

where λ_1 and λ_2 are Lagrange multipliers. Since the x_i's are a solution to the quadratic equation $2x(x-a) = \lambda_1 x + \lambda_2$, it follows that two of the x_i's have the same value. This implies that there exists either an optimal solution of the form $x_1 = x_2 = a - R$, $x_3 = a + 2R$, where R is a positive number satisfying

$$(a + 2R)(a - R)^2 = 1, \tag{4.1}$$

or an optimal solution with $x_1 = a - 2r$, $x_2 = x_3 = a + r$, where r is a positive number satisfying

$$(a - 2r)(a + r)^2 = 1. \tag{4.2}$$

The corresponding values for $(x_1 - a)^2 + (x_2 - a)^2 + (x_3 - a)^2$ are $6R^2$ and $6r^2$. From (4.1) and (4.2) we get

$$(a + 2R)(a - R)^2 - (a - 2r)(a + r)^2 = 0. \tag{4.3}$$

Equivalently, (4.3) can be written as

$$3(r^2 - R^2) + 2(r^3 + R^3) = 0. \tag{4.4}$$

But (4.4) implies that $r < R$. Hence we have the proof of the lemma for $n = 3$.

Next we prove the lemma for $n > 3$. Let $x = (x_1, x_2, \ldots, x_n)$ be an optimal solution for $B(n,a)$, where $n > 3$. Since for any permutation π of the set $\{1, \ldots, n\}$ $x_\pi = (x_{\pi(1)}, \ldots, x_{\pi(n)})$ is also an optimal solution, without loss of generality we may assume that $x_1 \leq x_2 \leq \cdots \leq x_n$. Suppose there exists i, j, and k such that $x_i < x_j \leq x_k$. Consider the three-dimensional minimization problem that results

when all the variables except the ith, jth, and kth stay fixed at the value of the corresponding component of x. Note that from the homogeneity of the constraint set this three-dimensional minimization can be reduced to the problem of computing $B(3, a')$ for some $a' \geq 1$ having an optimal solution which is a scalar multiple of (x_i, x_j, x_k). But this contradicts the correctness of the lemma for $n = 3$. □

COROLLARY 3. *$B(n, a) = a\xi$, where $\xi = \xi(n) \in [0, 1)$ satisfies*

$$a^n(1-\xi)^{n-1}(1+(n-1)\xi) = 1. \tag{4.5}$$

Proof. From Lemma 1 there exists an optimal solution satisfying

$$x_1 = \cdots = x_{n-1} = a(1-\xi)$$

for some $\xi \in [0, 1)$. Since the average value of the x_i's is a, we get $x_n = a(1+(n-1)\xi)$. Also, the product of the x_i's is 1. Hence, (4.5) holds. □

LEMMA 2. $B(n, a) \leq B(n-1, a) \leq \cdots \leq B(2, a) = \sqrt{a^2 - 1}$.

Proof. The fact that $B(2, a) = \sqrt{a^2 - 1}$ is immediate from Corollary 2. Next we prove the monotonicity of $B(n, a)$ in n. From (4.5) we have

$$n \ln a + (n-1)\ln(1-\xi) + \ln(1+(n-1)\xi) = 0. \tag{4.6}$$

Treating n as a continuous variable, and upon the implicit differentiation of (4.6), we obtain

$$\frac{d\xi}{dn} B + A = 0,$$

where

$$B = \frac{-n\xi}{(1+(n-1)\xi)(1-\xi)},$$

and

$$A = \ln(a(1-\xi)) + \frac{\xi}{(1+(n-1)\xi)}.$$

Since $B < 0$, it suffices to show that $A \leq 0$. By (4.6),

$$\ln(a(1-\xi)) = \frac{1}{n}\left(\ln(1-\xi) - \ln(1+(n-1)\xi)\right) = \frac{1}{n}\ln\left(\frac{(1-\xi)}{(1+(n-1)\xi)}\right). \tag{4.7}$$

From (4.7) we have

$$A = \frac{1}{n}\ln\left(1 - \frac{n\xi}{(1+(n-1)\xi)}\right) - \frac{\xi}{(1+(n-1)\xi)} = \frac{1}{n}\ln(1-nu) + u, \tag{4.8}$$

where $u = n\xi/[1+(n-1)\xi]$. Using the fact that $e^x \geq 1 + x$ for all x, from (4.8) it follows that $A \leq 0$, and hence we have the proof of monotonicity of $B(n, a)$. □

LEMMA 3. *Suppose that $A = X^{*^{-1}} C X^*$, where C is a circuit matrix with t equal to the size of the circuit, and $F(w) = \sum_{(i,j)\in E} a_{ij} e^{w_i - w_j}$. Then*

$$\frac{1}{t}\|\nabla F(w)\| \leq 2B\left(t, \frac{1}{t}F(w)\right).$$

Proof. Without loss of generality assume that the corresponding circuit in G_A is $\{(1,2),\ldots,(t-1,t)\}$. Thus

$$F(w) = e^{\Delta w_1 - \Delta w_2} + e^{\Delta w_2 - \Delta w_3} + \cdots + e^{\Delta w_t - \Delta w_1}, \tag{4.9}$$

where $w_i^* = \ln x_i^*$ and where $\Delta w_i = w_i - w_i^*$ for all $i = 1,\ldots,n$. From (4.9) it follows that

$$\nabla F(w) = (e^{\Delta w_1 - \Delta w_2} - e^{\Delta w_n - \Delta w_1}, \ldots, e^{\Delta w_t - \Delta w_1} - e^{\Delta w_{t-1} - \Delta w_t})^T. \tag{4.10}$$

Let $z = (z_1,\ldots,z_t)^T$, where

$$z_1 = e^{\Delta w_1 - \Delta w_2}, \quad z_2 = e^{\Delta w_2 - \Delta w_3}, \ldots, z_t = e^{\Delta w_t - \Delta w_1},$$

and let $z' = (z_2, z_3, \ldots, z_t, z_1)^T$. Then

$$\|\nabla F(w)\| = \|z - z'\|. \tag{4.11}$$

Let

$$a = \frac{1}{t} F(w), \quad \widehat{a} = (a, a, \ldots, a)^T \in \Re^t.$$

From (4.11) and the triangle inequality we get

$$\|\nabla F(w)\| \le \|z - \widehat{a}\| + \|z' - \widehat{a}\| = 2\|z - \widehat{a}\|. \tag{4.12}$$

Now from (4.12) and Lemma 1 we obtain

$$\frac{1}{t}\|\nabla F(w)\| \le \frac{2}{t}\|z - \widehat{a}\| \le \frac{2}{t}\sqrt{\frac{t}{t-1}}\|z - \widehat{a}\| = 2\sigma(z, \widehat{a}) \le 2B\left(t, \frac{1}{t}F(w)\right). \qquad \square$$

LEMMA 4. *Let β_k, $k = 1,\ldots,q$ be a set of positive numbers satisfying $\sum_{k=1}^q \beta_k = 1$. Let $\gamma \ge 1$ be any given number. For $x \in \Re^q$, let $H(x) = \sum_{k=1}^q \beta_k \sqrt{x_k^2 - 1}$. Then*

$$\max\left\{H(x) : \sum_{k=1}^q \beta_k x_k \le \gamma, \quad x_k \ge 1, \quad k = 1,\ldots,q\right\} = \sqrt{\gamma^2 - 1}.$$

Proof. Note that $H(x)$ is concave and the feasible region is convex. Thus, to prove the lemma it suffices to check that $x = (\gamma,\ldots,\gamma) \in \Re^q$ is a constrained stationary point. This can easily be established. $\square$

Proof of Theorem 3. Let $A = \sum_{k=1}^q \alpha_k A_k$, where the $A_k = X^{*^{-1}} C_k X^*$'s are the decomposition components ensured by Theorem 1 (iv). For each $k = 1,\ldots,q$, let E_k be the simple cycle in G_A corresponding to A_k, and let n_k be the size of the cycle. In particular, the positive entries of A_k are given by

$$a_{ij}^{(k)} = \alpha_k \frac{x_j^*}{x_i^*}, \quad (i,j) \in E_k.$$

Let $w^* = (\ln x_1^*, \ldots, \ln x_n^*)$ and $\Delta w_i = w_i - w_i^*$, $i = 1,\ldots,n$. For each $k = 1,\ldots,q$, define

$$F_k(w) = \sum_{(i,j)\in E_k} e^{\Delta w_i - \Delta w_j}. \tag{4.13}$$

Then we have

$$F(w) = \sum_{k=1}^{q} \alpha_k F_k(w) \tag{4.14}$$

and

$$F(w^*) = \sum_{k=1}^{q} \alpha_k n_k. \tag{4.15}$$

Suppose

$$G(w) - G(w^*) \leq \delta,$$

where $G(w) = \ln F(w)$. Equivalently, suppose that

$$\frac{F(w)}{F(w^*)} \leq e^{\delta}. \tag{4.16}$$

From (4.14) and (4.15) we obtain

$$\frac{F(w)}{F(w^*)} = \frac{\sum_{k=1}^{q} \alpha_k n_k \frac{F_k(w)}{n_k}}{\sum_{k=1}^{q} \alpha_k n_k}. \tag{4.17}$$

Define

$$\beta_k = \frac{\alpha_k n_k}{\sum_{k=1}^{q} \alpha_k n_k}, \quad k = 1, \ldots, q. \tag{4.18}$$

With this notation (4.16) can be written as

$$\sum_{k=1}^{q} \beta_k \frac{F_k(w)}{n_k} \leq e^{\delta}. \tag{4.19}$$

Since X^* simultaneously balances each A_k (see Corollary 1) we have

$$F_k(w) \geq n_k, \quad k = 1, \ldots, q. \tag{4.20}$$

From (4.17) it follows that

$$\nabla G(w) = \frac{\nabla F(w)}{F(w)} = \frac{\sum_{k=1}^{q} \alpha_k \nabla F_k(w)}{\sum_{k=1}^{q} \alpha_k F_k(w)}. \tag{4.21}$$

By (4.18), (4.20), and (4.21)

$$\|\nabla G(w)\| \leq \frac{\sum_{k=1}^{q} \alpha_k \|\nabla F_k(w)\|}{\sum_{k=1}^{q} \alpha_k n_k} = \sum_{k=1}^{q} \beta_k \frac{\|\nabla F_k(w)\|}{n_k}.$$

Hence from Lemma 3 we obtain

$$\|\nabla G(w)\| \leq 2 \sum_{k=1}^{q} \beta_k B\left(n_k, \frac{1}{n_k} F_k(w)\right). \tag{4.22}$$

Let

$$a_k = \frac{F_k(w)}{n_k}, \quad k = 1, \ldots, q. \tag{4.23}$$

Note that from (4.20) $a_k \geq 1$. From this, (4.22), and Lemma 2 we have

$$\|\nabla G(w)\| \leq 2 \sum_{k=1}^{q} \beta_k B(2, a_k) = 2 \sum_{k=1}^{q} \beta_k \sqrt{a_k^2 - 1}. \tag{4.24}$$

Since $\sum_{k=1}^{q} \beta_k a_k \leq e^{\delta}$ and $\sum_{k=1}^{q} \beta_k = 1$, from (4.24) and Lemma 4 we conclude that

$$\|\nabla G(w)\| \leq 2\sqrt{e^{2\delta} - 1}.$$

Hence, we have the proof of Theorem 3. ∎

5. Polynomial-time solvability of the matrix balancing problem. To complete the proof of polynomial-time solvability we also need the following result.

THEOREM 4. *Let A be an $n \times n$ nonnegative matrix, $a_{ii} = 0$, for all $i = 1, \ldots, n$, and G_A strongly connected. Let w^* be an exact minimizer of $G(w)$ in $\Re^n$. Given $\varepsilon \in (0, 1)$, define*

$$r = \frac{\varepsilon^2}{16\sqrt{2}}. \tag{5.1}$$

The condition

$$\|w - w^*\| \leq r \tag{5.2}$$

implies

$$G(w) - G(w^*) \leq \delta = \frac{\varepsilon^2}{16}. \tag{5.3}$$

Proof. From (5.2) it follows that

$$|\Delta w_i - \Delta w_j| \leq \sqrt{2} r, \tag{5.4}$$

where, as before, we use the notation $\Delta w_i = w_i - w_i^*$, $i = 1, \ldots, n$. Recall from the proof of Theorem 3 that

$$\frac{F(w)}{F(w^*)} = \sum_{k=1}^{q} \beta_k \frac{F_k(w)}{n_k}, \tag{5.5}$$

where $\sum_{k=1}^{q} \beta_k = 1$; see (4.17) and (4.18). Also recall that for each $k = 1, \ldots, q$ we have

$$F_k(w) = e^{\Delta w_{t_1} - \Delta w_{t_2}} + e^{\Delta w_{t_2} - \Delta w_{t_3}} + \cdots + e^{\Delta w_{t_{n_k - 1}} - \Delta w_{t_{n_k}}}, \tag{5.6}$$

where $\{t_1, \ldots, t_{n_k}\}$ is a subset of $\{1, \ldots, n\}$. Using (5.4) and (5.6) we get

$$F_k(w) \leq n_k e^{\sqrt{2} r},$$

which in view of (5.5) yields

$$\frac{F(w)}{F(w^*)} \le \sum_{k=1}^{q} \beta_k e^{\sqrt{2}r} = e^{\sqrt{2}r}.$$

Substituting (5.1) into the above inequality, we obtain (5.3). □

From Theorem 2, Theorem 3, and Corollary 2 it follows that in order to balance a given nonnegative $n \times n$ matrix A, whose diagonal entries are zero and G_A strongly connected, to the relative error of ε, it suffices to solve the convex program

$$\min\left\{G(w) = \ln \sum_{i,j=1}^{n} a_{ij} e^{w_i - w_j} : \|w\| \le R = \frac{\sqrt{n}(n-1)}{2} \ln \frac{e}{(n-1)v}\right\} \tag{5.7}$$

with the absolute accuracy $\delta = \varepsilon^2/16$. But the number of iterations of the ellipsoid method for solving any convex optimization problem of the form $\min\{G(w) : \|w\| \le R\}$ is bounded by $2n(n+1)\ln(R/r)$, provided that the set of ε solutions contains a ball of radius r (see, e.g., [4]). Hence by Theorem 4 (5.7) can be solved in

$$2n(n+1)\ln\frac{R}{r} = O\left(n^2 \ln\left(\frac{n}{\varepsilon}\ln\frac{1}{v}\right)\right) \tag{5.8}$$

iterations. Each iteration of the ellipsoid method requires $O(n^2)$ operations plus the overhead for evaluation of the function and its gradient. In our case the total complexity remains as $O(n^2)$ operations. These operations are to be performed over the numbers ω_i's having at most $O(\ln(\frac{n}{\varepsilon v}))$ digits before and after the decimal point. From Theorem 4, Theorem 3, and Corollary 2 we conclude that if $\|w - w^*\| \le r$, then $F(w)/F(w^*) \le e^{\delta}$. Thus, $\|w - w^*\| \le r$ implies that

$$\frac{\|\nabla F(w)\|}{e^{\delta} F(w^*)} \le \frac{\|\nabla F(w)\|}{F(w)} = \|\nabla G(w)\| \le \varepsilon.$$

Hence,

$$\|\nabla F(w)\| \le e^{\delta} F(w^*)\varepsilon \le eF(0)\varepsilon = e\sigma\varepsilon.$$

We thus have the following theorem.

THEOREM 5. *Let A be an $n \times n$ nonnegative matrix, $a_{ii} = 0$, for all $i = 1, \ldots, n$. Suppose that $G_A = (V, E)$ is strongly connected. Let $a_{\min} = \min\{a_{ij} : (i,j) \in E\}$, $\sigma = \sum_{(i,j)\in E} a_{ij}$, and $v = a_{\min}/\sigma$. For any given accuracy $\varepsilon \in (0,1)$, in $O\big(n^4 \ln\big(\frac{n}{\varepsilon}\ln\frac{1}{v}\big)\big)$ arithmetic operations over $O(\ln(\frac{n}{\varepsilon v}))$-bit numbers, we can compute a positive diagonal matrix $X = \operatorname{diag}(e^{w_1}, \ldots, e^{w_n})$ so that XAX^{-1} is balanced to the relative error of ε and the absolute error of $e\sigma\varepsilon$.* □

The convex programming problem (5.7) can also be solved via interior-point Newton methods in $O(m^{3.5}\sqrt{m}\ln(n/\epsilon\ln(1/v)))$ arithmetic operations (see [6]), where m is the number of positive entries of the input matrix A. However, for matrices with $m > n^{8/7}$ this complexity is inferior to the one stated in Theorem 5. It should be mentioned that these complexity bounds also apply to the doubly stochastic diagonal scaling of nonnegative matrices; see [5].

Acknowledgments. The authors wish to thank the anonymous referees for their helpful remarks.

REFERENCES

[1] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Studies in Applied Mathematics 15, SIAM, Philadelphia, PA, 1994.

[2] B. C. Eaves, A. J. Hoffman, U. G. Rothblum, and H. Schneider, *Line-sum-symmetric scaling of square nonnegative matrices*, Math. Programming Study, 25 (1985), pp. 124–141.

[3] J. Grad, *Matrix balancing*, Comput. J., 14 (1971), pp. 280–284.

[4] M. Grötschel, L. Lóvasz, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Springer-Verlag, Berlin, 1988.

[5] B. Kalantari and L. Khachiyan, *On the complexity of nonnegative matrix scaling*, Linear Algebra Appl., 240 (1996), pp. 87–103.

[6] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Studies in Applied Mathematics 13, SIAM, Philadelphia, PA, 1994.

[7] E. E. Osborne, *On pre-conditioning of matrices*, JACM, 7 (1960), pp. 338–345.

[8] U. G. Rothblum and H. Schneider, *Scaling of matrices which have prescribed row sums and column sums via optimization*, Linear Algebra Appl., 114/115 (1989), pp. 737–764.

[9] M. H. Schneider and S. A. Zenios, *A comparative study of algorithms for matrix balancing*, Oper. Res., 38 (1989), pp. 439–455.

[10] M. H. Schneider, *Matrix scaling, entropy minimization, and conjugate duality* (I): *Positivity conditions*, Linear Algebra Appl., 114/115 (1989), pp. 785–813.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, pp. 464–481, April 1997

015

CIRCULANT PRECONDITIONERS FOR MARKOV-MODULATED POISSON PROCESSES AND THEIR APPLICATIONS TO MANUFACTURING SYSTEMS*

WAI KI CHING†, RAYMOND H. CHAN‡, AND XUN YU ZHOU§

Abstract. The Markov-modulated Poisson process (MMPP) is a generalization of the Poisson process and is commonly used in modeling the input process of communication systems such as data traffic systems and ATM networks. In this paper, we give fast algorithms for solving queueing systems and manufacturing systems with MMPP inputs. We consider queueing systems where the input of the queues is a superposition of the MMPP which is still an MMPP. The generator matrices of these processes are tridiagonal block matrices with each diagonal block being a sum of tensor products of matrices. We are interested in finding the steady state probability distributions of these processes which are the normalized null vectors of their generator matrices. Classical iterative methods, such as the block Gauss–Seidel method, are usually employed to solve for the steady state probability distributions. They are easy to implement, but their convergence rates are slow in general. The number of iterations required for convergence increases like $O(m)$, where m is the size of the waiting spaces in the queues. Here, we propose to use the preconditioned conjugate gradient method. We construct our preconditioners by taking circulant approximations of the tensor blocks of the generator matrices. We show that the number of iterations required for convergence increases at most like $O(\log_2 m)$ for large m. Numerical results are given to illustrate the fast convergence.

As an application, we apply the MMPP to model unreliable manufacturing systems. The production process consists of multiple parallel machines which produce one type of product. Each machine has exponentially distributed up time, down time, and processing time for one unit of product. The interarrival of a demand is exponentially distributed and finite backlog is allowed. We consider hedging point policy as the production control. The average running cost of the system can be written in terms of the steady state probability distribution. Our numerical algorithm developed for the queueing systems can be applied to obtain the steady state distribution for the system and hence the optimal hedging point. Furthermore, our method can be generalized to handle the case when the machines have a more general type of repairing process distribution such as the Erlangian distribution.

Key words. Markov-modulated Poisson process, preconditioned conjugate gradient squared method, manufacturing systems, hedging point policy

AMS subject classifications. 65C20, 65F10

PII. S0895479895293442

1. Introduction.

The Markov-modulated Poisson process (MMPP) is a generalization of the Poisson process and is widely used as the input model of communication systems such as data traffic systems [8] and ATM networks [23]. An MMPP is a Poisson process whose instantaneous rate is itself a stationary random process which varies according to an irreducible n-state Markov chain. If n is 1, then the process is just a Poisson process. We say that the MMPP is in phase k, $1 \leq k \leq n$, when the underlying Markov process is in state k, and in this case the arrivals occur according

* Received by the editors October 20, 1995; accepted for publication (in revised form) by D. P. O'Leary June 11, 1996.

http://www.siam.org/journals/simax/18-2/29344.html

† Department of Applied Mathematics, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong and Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, Shatin, Hong Kong (wkching@se.cuhk.edu.hk).

‡ Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong (rchan@math.cuhk.edu.hk). This research was supported by HKRGC grant CUHK 316/94E.

§ Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, Shatin, Hong Kong (xyzhou@se.cuhk.edu.hk). This research was supported by RGC Earmarked grant CUHK 489/95E and RGC Earmarked grant CUHK 249/94E.

to a Poisson process of rate λ_k. The process is characterized by the generator matrix Q of the underlying Markov process and the rates $\lambda_1, \lambda_2, \ldots, \lambda_n$.

In this paper, we first discuss a numerical algorithm for solving the steady state probability distributions of queueing systems with MMPP inputs. We then relate queueing systems with MMPP inputs to the production process in unreliable manufacturing systems under the hedging point production control. Our algorithm can be applied to solve for the steady state probability distribution of these systems and hence their optimal hedging points.

We consider a queueing system with $(q+1)$ trunks, where each trunk has m waiting spaces and s multiple exponential servers. The analysis of these queueing systems can be used to determine call congestions in teletraffic networks with alternate routing; see Meier-Hellstern [13]. A call will overflow to other trunks if its first destination trunk is full and will be blocked from the system if all the trunks are full. The analysis of these queueing systems can be decomposed into the study of each trunk independently; see Meier-Hellstern [13]. For each trunk, the overflow from other trunks is modeled by a 2^q-state MMPP which is a superposition of q independent 2-state MMPPs; i.e., each trunk is an (MMPP/M/$s/s+m$) queue. The generator matrices of these processes are $(s+m+1)2^q \times (s+m+1)2^q$ tridiagonal block matrices with each diagonal block being a sum of tensor products of matrices. We are interested in finding the steady state probability distributions of the queues which are the normalized null vectors of the generator matrices.

Usually classical iterative methods, such as the block Gauss–Seidel method, are used to solve for the steady state probability distribution. They are easy to implement, but their convergence rates are slow in general; see the numerical results in section 7. Here, we propose to use the preconditioned conjugate gradient (PCG) method. Our preconditioners are constructed by taking circulant approximations of the tensor blocks of the generator matrix. We prove that the preconditioned system has singular values clustered around 1 independent of the size of the waiting spaces m. Hence the conjugate gradient method will converge very fast when employed to solve the preconditioned system for large m. In fact, we prove that the number of iterations required for convergence grows at most like $O(\log_2 m)$. Numerical examples are given in section 7 to illustrate the fast convergence. For the case of a single server ($s = 1$), our generator matrix corresponds to a class of quasi-birth–death (QBD) processes which can be solved efficiently by the folding algorithm; see Ye and Li [22]. We will compare the complexity of our PCG method with that of the folding algorithm in section 7. The cost of our PCG method increases more slowly than that of the folding algorithm when the problem size increases. In fact, for large values of q ($q \geq 6$), the PCG-type method is more efficient than the folding algorithm.

The analysis of the MMPP queueing systems can be applied to the production planning of manufacturing systems. We consider manufacturing systems of multiple parallel machines producing one type of product. Usually positive inventory is stored to hedge against uncertain situations such as the breakdown of machines and the shortfall of products; see Akella and Kumar [1]. It is well known that the hedging point policy is optimal for one-machine manufacturing systems in some simple situations; see [1, 3, 10, 11]. For two-machine flowshops, hedging policies are no longer optimal but near optimal; see [16, 15]. A hedging point policy is characterized by a number h: the machines keep producing the product at the maximum possible production rate if the inventory level is less than h, maintain the inventory level h as far as they can if the inventory level reaches h, and stop producing if the inventory level exceeds h.

When the optimal policy is a zero-inventory policy (i.e., the hedging point is zero), then the policy matches with the just-in-time (JIT) policy. The JIT policies have strongly been favored in real-life production systems for process discipline reasons even when they are not optimal. By using the JIT policy, the Toyota company has managed to reduce work-in-process and cycle time in the presence of the stochastic situations mentioned above; see Monden [14]. We focus on finding optimal hedging point policies for the manufacturing systems.

We note that in [1, 3, 10, 11] only one-machine systems are considered, and, in addition, the repairing process of the machine is assumed to be exponentially distributed. Ching and Zhou [6] consider one-machine manufacturing systems with the repairing process being Erlangian distributed. The algorithm proposed here can deal with the more general case of multiple machines. Each machine is unreliable and has exponential up time and down time, and the demand is a Poisson process. The production process of the machines can be modeled as an MMPP. The generator matrix for the machine-inventory system is a particular case of the queueing systems discussed above, with the queue size m being the size of the inventory which in practice can easily go up to the thousands. Our numerical method developed for the queueing networks above is well suited for solving the steady state probability distribution for these processes. Given a hedging point, the average running cost of the machine-inventory system can be written in terms of the steady state probability distribution. Hence the optimal hedging point can also be obtained. Moreover, our algorithm can also handle the case when the repair time has a more general distribution, e.g., the Erlangian distribution.

The outline of the paper is as follows. In section 2, we present the generator matrix for the queueing system (MMPP/M/s/$s+m$). In section 3, we construct preconditioners by taking circulant approximations of the tensor blocks of the generator matrices. In section 4, we prove that the preconditioned systems have singular values clustered around 1. The cost count of our method is given in section 5. In section 6, we apply our method to the production planning of manufacturing systems with multiple parallel machines. Numerical examples are given in section 7 to illustrate the fast convergence rate of our method. Finally, concluding remarks are given in section 8.

2. The queueing system. In this section, we present the queueing system (MMPP/M/s/$s+m$) arising in telecommunication networks; see, for instance, Meier-Hellstern [13]. In order to construct the generator matrix of the queueing process, we first define the following queueing parameters:

(i) $1/\lambda$, the mean arrival time of the exogenously originating calls,
(ii) $1/\mu$, the mean service time of each server,
(iii) s, the number of servers,
(iv) m, the number of waiting spaces in the queue,
(v) q, the number of overflow queues, and
(vi) (Q_j, Λ_j), $1 \le j \le q$, the parameters of the MMPP's modeling overflow parcels, where

$$Q_j = \begin{pmatrix} \sigma_{j1} & -\sigma_{j2} \\ -\sigma_{j1} & \sigma_{j2} \end{pmatrix} \quad \text{and} \quad \Lambda_j = \begin{pmatrix} \lambda_j & 0 \\ 0 & 0 \end{pmatrix}. \tag{1}$$

Here σ_{j1}, σ_{j2}, and λ_j, $1 \le j \le q$, are positive MMPP parameters. Conventionally, an infinitesimal generator Q has nonnegative off-diagonal entries and zero row sums.

For ease of presentation, in our discussion all the infinitesimal generators are of the form $-Q^t$, which has nonpositive off-diagonal entries and zero column sums.

The input of the queue comes from the superposition of several independent MMPPs, which is still an MMPP and is parametrized by two $2^q \times 2^q$ matrices (Q, Γ). Here

$$Q = (Q_1 \otimes I_2 \otimes \cdots \otimes I_2) + (I_2 \otimes Q_2 \otimes I_2 \otimes \cdots \otimes I_2) + \cdots + (I_2 \otimes \cdots \otimes I_2 \otimes Q_q), \tag{2}$$

$$\Lambda = (\Lambda_1 \otimes I_2 \otimes \cdots \otimes I_2) + (I_2 \otimes \Lambda_2 \otimes I_2 \otimes \cdots \otimes I_2) + \cdots + (I_2 \otimes \cdots \otimes I_2 \otimes \Lambda_q), \tag{3}$$

and

$$\Gamma = \Lambda + \lambda I_{2^q},$$

where I_2 and I_{2^q} are the 2×2 and $2^q \times 2^q$ identity matrices, respectively, and $\otimes$ denotes the Kronecker tensor product. In the following, we will drop the subscript of the identity matrix I if the dimension of the matrix is clear from the context.

We can regard our (MMPP/M/s/$s+m$) queue as a Markov process on the state space

$$\{(i, j) \mid 0 \le i \le s + m, 1 \le j \le 2^q\}.$$

The number i corresponds to the number of calls at the destination, while j corresponds to the state of the Markov process with generator matrix Q. Hence the generator matrix of the queueing process is given by the following $(s+m+1)2^q \times (s+m+1)2^q$ tridiagonal block matrix A:

$$A = \begin{pmatrix} Q+\Gamma & -\mu I & & & & & & 0 \\ -\Gamma & Q+\Gamma+\mu I & -2\mu I & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & -\Gamma & Q+\Gamma+s\mu I & -s\mu I & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & -\Gamma & Q+\Gamma+s\mu I & -s\mu I & \\ 0 & & & & & -\Gamma & Q+s\mu I & \end{pmatrix}. \tag{4}$$

For simplicity, let us write $n = (s + m + 1)2^q$. The steady state probability distribution vector $\mathbf{p} = (p_1, p_2, \ldots, p_n)^t$ is the solution to the matrix equation $A\mathbf{p} = \mathbf{0}$ with constraints

$$\sum_{i=1}^{n} p_i = 1$$

and

$$p_i \ge 0 \quad \text{for all } 1 \le i \le n.$$

Note that the matrix A is irreducible and has zero column sums, positive diagonal entries, and nonpositive off-diagonal entries. From Perron–Frobenius theory, the matrix A has a one-dimensional null space with a positive null vector; see Varga [20, p. 30]. Therefore, the steady state probability distribution vector $\mathbf{p}$ exists.

Many useful quantities such as the steady state distribution of the number of calls at the destination, the blocking probability, and the waiting time distribution can be obtained from the vector $\mathbf{p}$; see Meier-Hellstern [13]. We note that $\mathbf{p}$ can be obtained by normalizing the solution $\mathbf{x}$ of the nonsingular system

$$G\mathbf{x} \equiv (A + \mathbf{e}_n\mathbf{e}_n^t)\mathbf{x} = \mathbf{e}_n. \tag{5}$$

Here $\mathbf{e}_n = (0, \ldots, 0, 1)^t$ is an n-vector. The matrix G is nonsingular because it is irreducible diagonally dominant with the last column being strictly diagonally dominant. We will solve the linear system (5) by conjugate gradient (CG)-type methods; see [2, 18]. The convergence rate of CG-type methods depends on the distribution of the singular values of the matrix G. The more clustered the singular values of G are, the faster the convergence rate will be; see Axelsson and Barker [2].

However, this is not the case for our matrix G, and we will see in the numerical results in section 7 that the convergence for system (5) is very slow. To speed up the convergence, a preconditioner is used. In essence, we solve instead of (5) the preconditioned system

$$GC^{-1}\mathbf{w} = \mathbf{e}_n \tag{6}$$

for $\mathbf{w}$ by CG-type methods. Obviously, the solution $\mathbf{x}$ to (5) is given by $C^{-1}\mathbf{w}$. A good preconditioner C is an easy-to-construct matrix, the preconditioned matrix GC^{-1} has singular values clustered around 1, and the preconditioned system $C\mathbf{y} = \mathbf{r}$ can be solved easily for any vector $\mathbf{r}$; see Axelsson and Barker [2]. We will show that our preconditioner satisfies these three criteria in the next three sections.

3. Construction of our preconditioners. In this section, we discuss the construction of preconditioners for the linear system (6). Our preconditioner C is constructed by exploiting the block structure of the generator matrix A in (4). Notice that the generator A can be written as the sum of tensor products:

$$A = I \otimes Q + B \otimes I + R \otimes \Lambda, \tag{7}$$

where B and R are $(s + m + 1) \times (s + m + 1)$ matrices given by

$$B = \begin{pmatrix} \lambda & -\mu & & & & & 0 \\ -\lambda & \lambda+\mu & -2\mu & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & -\lambda & \lambda+s\mu & -s\mu & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & -\lambda & \lambda+s\mu & -s\mu \\ 0 & & & & & -\lambda & s\mu \end{pmatrix}$$

and

$$R = \begin{pmatrix} 1 & & & & 0 \\ -1 & 1 & & & \\ & -1 & \ddots & & \\ & & \ddots & 1 & \\ 0 & & & -1 & 0 \end{pmatrix}.$$

For small s, we observe that B and R are close to the tridiagonal Toeplitz matrices

$$\text{tridiag}[-\lambda, \lambda + s\mu, -s\mu] \qquad \text{and} \qquad \text{tridiag}[-1, 1, 0],$$

respectively. Our preconditioner is then obtained by taking the "circulant approximation" of the matrices B and R, which are defined by $c(B)$ and $c(R)$ as follows:

$$c(B) = \begin{pmatrix} \lambda + s\mu & -s\mu & & & -\lambda \\ -\lambda & \lambda + s\mu & -s\mu & & \\ & \ddots & \ddots & \ddots & \\ & & -\lambda & \lambda + s\mu & -s\mu \\ -s\mu & & & -\lambda & \lambda + s\mu \end{pmatrix} \tag{8}$$

and

$$c(R) = \begin{pmatrix} 1 & & & & -1 \\ -1 & 1 & & & \\ & -1 & \ddots & & \\ & & \ddots & 1 & \\ 0 & & & -1 & 1 \end{pmatrix}. \tag{9}$$

We note that $c(B)$ and $c(R)$ are Strang's circulant approximations of the Toeplitz matrices tridiag$[-\lambda, \lambda + s\mu, -s\mu]$ and tridiag$[-1, 1, 0]$, respectively; see Chan [5]. Clearly, we have the following lemma.

LEMMA 1. $\text{rank}(B - c(B)) = s + 1$ *and* $\text{rank}(R - c(R)) = 1$.

Using the theory of circulant matrices (see Davis [7]) we also have the following.

LEMMA 2. *The matrices* $c(B)$ *and* $c(R)$ *can be diagonalized by the discrete Fourier transform matrix* F; *i.e.,*

$$F^*(c(B))F = \Phi \quad \text{and} \quad F^*(c(R))F = \Psi,$$

where both Φ *and* Ψ *are diagonal matrices. The eigenvalues of* $c(B)$ *and* $c(R)$ *are given by*

$$\phi_j = \lambda(1 - e^{\frac{-2\pi i(j-1)}{s+m+1}}) + s\mu(1 - e^{\frac{-2\pi i(j-1)(s+m)}{s+m+1}}), \quad j = 1, \ldots, s + m + 1 \tag{10}$$

and

$$\psi_j = 1 - e^{\frac{-2\pi i(j-1)}{s+m+1}}, \quad j = 1, \ldots, s + m + 1. \tag{11}$$

Thus, the matrices $c(B)$ and $c(R)$ can be inverted easily by using fast Fourier transforms.

We first approximate our matrix A in (7) (and hence G in (5)) by

$$D = I \otimes Q + c(B) \otimes I + c(R) \otimes \Lambda. \tag{12}$$

We observe that D is irreducible and has zero column sums, positive diagonal entries, and nonpositive off-diagonal entries. Hence D is singular and has a null space of dimension one. Moreover, D is unitarily similar to a diagonal block matrix:

$$(F^* \otimes I)D(F \otimes I) = I \otimes Q + \Phi \otimes I + \Psi \otimes \Lambda = \text{diag}(D_1, C_2, C_3, \ldots, C_{s+m+1}). \tag{13}$$

Here the blocks are

$$C_i = Q + \phi_i I + \psi_i \Lambda, \qquad i = 2, \ldots, s+m+1, \tag{14}$$

and $D_1 = Q$ with D_1 being the only singular block.

Let

$$C_1 = Q + \mathbf{e}_{2^q}\mathbf{e}_{2^q}^t, \tag{15}$$

where $\mathbf{e}_{2^q} = (0, \ldots, 0, 1)^t$ is a 2^q-vector. Since C_1 is irreducible diagonally dominant with the last column being strictly diagonally dominant, it is nonsingular. Our preconditioner C for the matrix G in (5) is defined as

$$C = (F \otimes I)\mathrm{diag}(C_1, C_2, \ldots, C_{s+m+1})(F^* \otimes I), \tag{16}$$

which is clearly nonsingular.

4. Convergence analysis. In this section, we study the convergence rate of our algorithm when m, the number of waiting spaces, is large. In the queueing systems considered in Meier-Hellstern [13], the number of waiting spaces m in each queue is much larger than the number of overflow queues q. In section 6, we apply the MMPP to model manufacturing systems of q parallel machines and m possible inventory states. In practice, the number of possible inventory states is much larger than the number of machines in the manufacturing systems and can easily go up to thousands.

We prove that if the queueing parameters λ, μ, s, q, and σ_{ij} are fixed independent of m, then the preconditioned system GC^{-1} in (6) has singular values clustered around 1 as m tends to infinity. Hence when CG-type methods are applied to solve the preconditioned system (6), we expect fast convergence. Numerical examples are given in section 7 to demonstrate our claim. We start the proof by the following lemma.

LEMMA 3. *We have* $\mathrm{rank}(G - C) \le (s+2)2^q + 2$.

Proof. We note that by (5) we have $\mathrm{rank}(G - A) = 1$. From (7), (12), and Lemma 1, we see that $\mathrm{rank}(A - D) = (s+2)2^q$. From (13), (15), and (16), we see that D and C differ by a rank-one matrix. Therefore, we have

$$\mathrm{rank}(G - C) \le \mathrm{rank}(G - A) + \mathrm{rank}(A - D) + \mathrm{rank}(D - C) = (s+2)2^q + 2.$$

Hence the inequality is proved. ∎

THEOREM 1. *The preconditioned matrix* GC^{-1} *has at most* $2((s+2)2^q + 2)$ *singular values not equal to* 1.

Proof. We first note that

$$GC^{-1} = I + (G - C)C^{-1} \equiv I + L_1,$$

where $\mathrm{rank}(L_1) \le (s+2)2^q + 2$ by Lemma 3. Therefore,

$$C^{-*}G^*GC^{-1} - I = L_1^*(I + L_1) + L_1$$

is a matrix of rank at most $2((s+2)2^q + 2)$. ∎

Thus the number of singular values of GC^{-1} that are distinct from 1 is a constant independent of m. In order to show fast convergence of PCG-type methods with preconditioner C, one still needs an estimate of $\sigma_{\min}(GC^{-1})$, the smallest singular value of GC^{-1}. If $\sigma_{\min}(GC^{-1})$ is uniformly bounded away from zero independent of m, then the method converges in $O(1)$ iterations; if $\sigma_{\min}(GC^{-1})$ decreases like

$O(m^{-\alpha})$ for some $\alpha > 0$, then the method converges in at most $O(\log_2 m)$ steps; see Van der Vorst [19] or Chan [4, Lemma 3.8.1].

In the remainder of this section, we show that even in the worst case in which $\sigma_{\min}(GC^{-1})$ decreases in an order faster than $O(m^{-\alpha})$ for any $\alpha > 0$ (e.g., like $O(e^{-m})$), we can still have a fast convergence rate. Note that in this case the matrix equation (6) is very ill conditioned. Our trick is to consider a regularized equation of (6) as follows:

$$C^{-*}(G^*G + m^{-4-\beta}I)C^{-1}\mathbf{w} = C^{-*}G^*\mathbf{e}_n, \tag{17}$$

where β is any positive constant.

In the following, we prove that the regularized preconditioned matrix

$$C^{-*}(G^*G + m^{-4-\beta}I)C^{-1}$$

has eigenvalues clustered around 1 and its smallest eigenvalues decrease at a rate no faster than $O(m^{-4-\beta})$. Hence PCG-type methods will converge in at most $O(\log_2 m)$ steps when applied to solve the preconditioned linear system (17). Moreover, we prove that the 2-norm of the error introduced by the regularization tends to zero at a rate of $O(m^{-\beta})$. To prove our claim, we must get an estimate of the upper and lower bounds for $||C^{-1}||_2$. We begin our proof by the following lemma.

LEMMA 4. *Given any matrix W, if the smallest eigenvalue of $W + W^*$, denoted by $\lambda_{\min}(W + W^*)$, satisfies $\lambda_{\min}(W + W^*) \geq \delta > 0$, then $||W^{-1}||_2 \leq 2/\delta$.*

Proof. For any arbitrary $\mathbf{x}$, using the Cauchy–Schwarz inequality, we have

$$\delta||\mathbf{x}||_2^2 \leq \lambda_{\min}(W + W^*)||\mathbf{x}||_2^2 \leq \mathbf{x}^*(W + W^*)\mathbf{x} = 2\mathbf{x}^*W\mathbf{x} \leq 2||\mathbf{x}||_2||W\mathbf{x}||_2.$$

Since $W\mathbf{x}$ is arbitrary, this implies $||W^{-1}||_2 \leq 2/\delta$. ∎

Now we are ready to estimate $||C^{-1}||_2$.

LEMMA 5. *Let the queueing parameters λ, μ, s, q, and σ_{ij} be independent of m. Then there exist positive constants τ_1 and τ_2 independent of m such that*

$$\tau_1 \leq ||C^{-1}||_2 \leq \tau_2 m^2.$$

Proof. We first prove the left-hand side of the inequality. From (16), we see that C is unitarily similar to a diagonal block matrix. We therefore have

$$||C||_2 = \max\left\{||C_1||_2, ||C_2||_2, \ldots, ||C_{s+m+1}||_2\right\}.$$

Using (14), (10), and (11), it is straightforward to check that $||C_i||_1$ and $||C_i||_\infty$, $1 \leq i \leq s + m + 1$, are all bounded above by

$$\frac{1}{\tau_1} \equiv q\left(\max_j\{\sigma_{j1}\} + \max_j\{\sigma_{j2}\}\right) + 2(\lambda + s\mu + 1).$$

Using the inequality

$$||\cdot||_2 \leq \sqrt{||\cdot||_1||\cdot||_\infty},$$

we see that $||C_i||_2$, $i = 1, \ldots, s + m + 1$, are all bounded above by $1/\tau_1$. Thus $||C||_2 \leq 1/\tau_1$ and hence $\tau_1 \leq ||C^{-1}||_2$.

Next we prove the right-hand side of the inequality. We note again by (16) that

$$||C^{-1}||_2 = \max\left\{||C_1^{-1}||_2, ||C_2^{-1}||_2, \ldots, ||C_{s+m+1}^{-1}||_2\right\}.$$

From (15), we can see that C_1 is a $2^q \times 2^q$ nonsingular matrix with entries independent of m. Thus $||C_1^{-1}||_2$ is bounded independent of m. To obtain bounds for $||C_i^{-1}||_2$, $i = 2, \ldots, s+m+1$, we first symmetrize the matrices. Define $\Sigma = \Sigma_1 \otimes \cdots \otimes \Sigma_q$, where

$$\Sigma_j = \begin{pmatrix} 1 & 0 \\ 0 & \frac{\sigma_{j1}}{\sigma_{j2}} \end{pmatrix}, \qquad j = 1, \ldots, q.$$

We see that $||\Sigma||_2$ and $||\Sigma^{-1}||_2$ are bounded independent of m. By (1) and (2), we see that $Q\Sigma$ is a symmetric semidefinite matrix. Thus

$$C_i\Sigma = Q\Sigma + \phi_i\Sigma + \psi_i\Lambda\Sigma, \qquad i = 2, \ldots, s+m+1,$$

are symmetric matrices too. By (11), we see that $(\psi_i\Lambda\Sigma + (\psi_i\Lambda\Sigma)^*)$, $i = 2, \ldots, s+m+1$, are diagonal positive semidefinite matrices. Therefore,

$$\lambda_{\min}(C_i\Sigma + (C_i\Sigma)^*) \geq \lambda_{\min}(\phi_i\Sigma + (\phi_i\Sigma)^*), \qquad i = 2, \ldots, s+m+1. \tag{18}$$

From (10), we have

$$\lambda_{\min}(\phi_i\Sigma + (\phi_i\Sigma)^*) \geq \lambda ||\Sigma^{-1}||_2^{-1} \sin^2\left(\frac{(i-1)\pi}{s+m+1}\right), \qquad i = 2, \ldots, s+m+1.$$

Since

$$\sin\theta \geq \min\left\{\frac{2\theta}{\pi}, 2\left(1 - \frac{\theta}{\pi}\right)\right\} \quad \forall \theta \in [0, \pi],$$

we have

$$\begin{aligned}\lambda_{\min}(\phi_i\Sigma + (\phi_i\Sigma)^*) &\geq \lambda ||\Sigma^{-1}||_2^{-1} \min\left\{\frac{4(i-1)^2}{(s+m+1)^2}, 4\left(1 - \frac{i-1}{s+m+1}\right)^2\right\} \\ &\geq \frac{4\lambda}{m^2} ||\Sigma^{-1}||_2^{-1}, \qquad i = 2, \ldots, s+m+1.\end{aligned}$$

By Weyl's theorem [9, p. 181], we then have

$$\lambda_{\min}(\phi_i\Sigma + (\phi_i\Sigma)^*) \geq \frac{\tau}{m^2}, \quad i = 2, \ldots, s+m+1,$$

where $\tau = 4\lambda ||\Sigma^{-1}||_2^{-1}$ is a positive constant independent of m.

Thus by (18) we get

$$\lambda_{\min}(C_i\Sigma + (C_i\Sigma)^*) \geq \frac{\tau}{m^2}, \quad i = 2, \ldots, s+m+1.$$

Hence by Lemma 4 we have

$$||\Sigma^{-1}C_i^{-1}||_2 \leq \frac{2}{\tau} m^2, \qquad i = 2, \ldots, s+m+1.$$

Therefore,

$$||C_i^{-1}||_2 \leq ||\Sigma||_2 ||\Sigma^{-1}C_i^{-1}||_2 \leq \frac{2m^2}{\tau} ||\Sigma||_2, \qquad i = 2, \ldots, s+m+1.$$

Since $||C_1^{-1}||_2$ is bounded above independent of m, we have

$$||C^{-1}||_2 \le \max\left\{||C_1^{-1}||_2, \frac{2m^2}{\tau}||\Sigma||_2\right\} \equiv \tau_2 m^2,$$

where τ_2 is a positive constant independent of m. Hence we have proved the lemma. □

THEOREM 2. *Let the queueing parameters λ, μ, s, q, and σ_{ij} be independent of m. Then for any positive β the regularized preconditioned matrix*

$$C^{-*}(G^*G + m^{-4-\beta}I)C^{-1} \tag{19}$$

has eigenvalues clustered around 1 *and the smallest eigenvalue decreases at a rate no faster than $O(m^{-4-\beta})$. Furthermore, the error introduced by the regularization is of the order $O(m^{-\beta})$.*

Proof. We note by Theorem 1 that

$$C^{-*}(G^*G + m^{-4-\beta}I)C^{-1} = I + L_2 + m^{-4-\beta}C^{-*}C^{-1},$$

where L_2 is a Hermitian matrix with $\text{rank}(L_2) \le 2((s+2)2^q + 2)$. By Lemma 5, we have

$$\lim_{m\to\infty} m^{-4-\beta}||C^{-*}C^{-1}||_2 \le \lim_{m\to\infty} m^{-\beta} = 0.$$

Thus by Cauchy's interlace theorem [9, p. 184] the regularized preconditioned matrix in (19) has eigenvalues clustered around 1 as m tends to infinity. The error introduced by the regularization is given by $m^{-4-\beta}||C^{-*}C^{-1}||_2$, which by Lemma 5 tends to zero like $O(m^{-\beta})$.

As for the smallest eigenvalue of the regularized preconditioned matrix in (19), we note that

$$\min_{||\mathbf{x}||_2=1} \frac{\mathbf{x}^*(G^*G + m^{-4-\beta})\mathbf{x}}{\mathbf{x}^*C^*C\mathbf{x}} \ge \frac{\min_{||\mathbf{x}||_2=1} \mathbf{x}^*(G^*G + m^{-4-\beta})\mathbf{x}}{\max_{||\mathbf{x}||_2=1} \mathbf{x}^*C^*C\mathbf{x}} \ge \frac{\tau_1}{m^{4+\beta}}, \tag{20}$$

where the rightmost inequality follows from Lemma 5. We recall that τ_1 and β are positive constants independent of m. Hence the smallest eigenvalue of the regularized preconditioned matrix in (19) decreases no faster than $O(m^{-4-\beta})$. □

Thus we conclude that PCG-type methods applied to (17) with $\beta > 0$ will converge in at most $O(\log_2 m)$ steps; see Van der Vorst [19] or Chan [4, Lemma 3.8.1]. To minimize the error introduced by the regularization, one can choose a large β. Recall that regularization is required only when the smallest singular value of the matrix GC^{-1} in (6) tends to zero faster than $O(m^{-\alpha})$ for any $\alpha > 0$. In view of Lemma 5 (or cf. (20)), this can happen only when the smallest singular value of G has the same decaying rate. This will imply that the matrix G is very ill conditioned. We note, however, that in all our numerical tests in section 7 we found that there is no need to add the regularization.

5. Cost analysis. In this section, we derive the computational cost of the PCG-type method. We compare our PCG method with the block Gauss–Seidel (BGS) method used in Meier-Hellstern [13] and the folding algorithm of Ye and Li [22]. We show that the cost for PCG-type algorithms is $O(2^q(s+m+1)\log_2(s+m+1) + q(s+m+1)2^q)$. The computational cost per iteration of the BGS method is

$O((s+m+1)2^{2q})$; see Meier-Hellstern [13]. Thus PCG-type methods require an extra $O(\log_2(s+m+1))$ of work per iteration compared with the BGS method. However, as we will soon see in the numerical examples of section 7, the fast convergence of our method can more than compensate for this minor overhead in each iteration.

When the queue has a single server, i.e., $s = 1$, our generator matrix A corresponds to a class of QBD processes which can be solved efficiently by the folding algorithm of Ye and Li [22]. The complexity of the folding algorithm is approximately $\frac{26\alpha}{3}2^{3q}\log_2(s+m+1)+3(s+m+1)2^{2q}$ operations, where $1 \le \alpha < 2$. We will compare the computational cost of our PCG method with the folding algorithm in section 7. Our PCG method is more efficient than the folding algorithm for large problems.

In PCG-type algorithms for (6), the main cost per iteration is to compute the matrix-vector multiplication of the form $GC^{-1}\mathbf{y}$ twice for some vector $\mathbf{y}$. By using the block tensor structure of A in (7), the multiplication of $G\mathbf{z}$ requires $(s+m+1)q2^q$ operations for any vector $\mathbf{z}$. By (16), we see that $C^{-1}\mathbf{y}$ is given by

$$(F \otimes I)\mathrm{diag}(C_1^{-1}, C_2^{-1}, \ldots, C_{s+m+1}^{-1})(F^* \otimes I)\mathbf{y}.$$

It involves the matrix-vector multiplications of the form

$$(F^* \otimes I)\mathbf{z} \qquad \text{and} \qquad (F \otimes I)\mathbf{z}.$$

By using fast Fourier transforms, they can be obtained in $6(s+m+1)2^q\log_2(s+m+1)$ operations. The vector

$$\mathrm{diag}(C_1^{-1}, C_2^{-1}, \ldots, C_{s+m+1}^{-1})\mathbf{z}$$

can be obtained by solving $(s+m+1)$ linear systems involving the matrices C_i, $i = 1, \ldots, s+m+1$. Since each matrix is of size $2^q \times 2^q$, if GE is used, $O((s+m+1)2^{3q})$ operations will be required. We now show that it can be reduced to $O((s+m+1)q2^q)$ operations.

First we recall from the definitions of C_i, Q, and Λ in (14), (2), and (3) that

$$\begin{aligned} C_i = {} & ((Q_1 + \phi_i I + \psi_i \Lambda_1) \otimes I \otimes \cdots \otimes I) + (I \otimes (Q_2 + \phi_i I + \psi_i \Lambda_2) \otimes I \otimes \cdots \otimes I) \\ & + \cdots + (I \otimes I \otimes \cdots \otimes (Q_q + \phi_i I + \psi_i \Lambda_q)), \end{aligned} \tag{21}$$

where Q_j and Λ_j, $j = 1, \ldots, q$, are given in (1). By using Schur's triangularization theorem [9, p. 79], we can find 2×2 unitary matrices U_{ij} and lower triangular matrices L_{ij} such that

$$U_{ij}^*(Q_j + \phi_i I + \psi_i \Lambda_j)U_{ij} = L_{ij}, \qquad 1 \le i \le s+m+1, \quad 1 \le j \le q. \tag{22}$$

For $i = 1, \ldots, s+m+1$, define

$$U_i \equiv U_{i1} \otimes \cdots \otimes U_{iq}$$

and

$$L_i \equiv (L_{i1} \otimes I \otimes \cdots \otimes I) + (I \otimes L_{i2} \otimes I \otimes \cdots \otimes I) + \cdots + (I \otimes I \otimes \cdots \otimes I \otimes L_{iq}).$$

We see from (21) and (22) that

$$U_i^* C_i U_i = L_i, \qquad 1 \le i \le s+m+1.$$

Hence the vector $C_i^{-1}\mathbf{w}$ can be computed as $U_i L_i^{-1} U_i^* \mathbf{w}$.

The matrix-vector multiplication of the form $U_j\mathbf{w}$ and $U_j^*\mathbf{w}$ can be done in $2(q2^q)$ operations by making use of the formula

$$U_j\mathbf{w} = (U_{1j} \otimes I \otimes \cdots \otimes I)(I \otimes U_{2j} \otimes I \otimes \cdots \otimes I) \cdots (I \otimes I \otimes \cdots \otimes I \otimes U_{qj})\mathbf{w}.$$

We note that the matrix L_i is a lower triangular matrix and each row of it has at most q nonzero entries. Hence $L_i^{-1}\mathbf{w}$ can be obtained in $q2^q$ operations. Thus for any vector $\mathbf{w}$ the vector $C_i^{-1}\mathbf{w}$ can be obtained in $3(q2^q)$ operations. Hence we conclude that the vector

$$\operatorname{diag}(C_1^{-1}, C_2^{-1}, \ldots, C_{s+m+1}^{-1})\mathbf{r}$$

can be computed in approximately $3(s+m+1)q2^q$ operations.

In summary, each iteration of PCG-type methods needs $2(6(s+m+1)2^q \log_2(s+m+1)+4(s+m+1)q2^q) \approx O(m \log_2 m)$ operations, as compared to $O((s+m+1)2^{2q}) \approx O(m)$ operations required by the BGS method. As we proved in section 4, PCG-type methods will converge in at most $O(\log_2 m)$ steps (see also the numerical results in section 7); therefore, the total complexity of our methods will be $O(m \log_2^2 m)$. As a comparison, the numerical results in section 7 show that the number of iterations required for convergence for the BGS method increases linearly like $O(m)$. Therefore, the total complexity of the BGS method is about $O(m^2)$ operations.

As for storage, PCG-type methods, the folding algorithm, and the BGS method require $O(2^q(s+m+1))$ memory. Clearly, at least $O(2^q(s+m+1))$ memory is required to store the approximated solution in each iteration.

6. The failure-prone manufacturing systems. In this section, we study a general kind of failure-prone manufacturing system. These systems consist of q multiple parallel machines producing one type of product. Each machine is subject to random breakdowns and repairs. The processing time for one unit of product, the up time, and the down time of each machine are exponentially distributed. The interarrival time of a demand is exponentially distributed. The systems allow finite backlog and a penalty cost is associated with the rejection of a demand. Moreover, there is an inventory cost for holding each unit of product and a shortfall cost for each unit of backlog.

The hedging point policy has been shown to be optimal for one-machine one-product manufacturing systems with repair time exponentially distributed; see [1, 3, 10, 11]. In those works, the discrete inventory levels of the product are approximated by a continuous fluid flow model. Analytic optimal control is found to be threshold (hedging point)-type by solving a pair of Hamilton–Jacobi–Bellman equations. The control is optimal in the sense that it minimizes the average (or discounted) running cost of the manufacturing systems. In this paper, we focus on finding the optimal hedging point for the manufacturing systems under consideration.

It should be noted that in [1, 3, 10, 11, 21, 17] only one machine is considered and the machine has only two states—up and down. Here we consider q parallel unreliable machines. The production process of the machines is then an MMPP. The states of the machines and the inventory level can be modeled as an irreducible continuous time Markov chain. For different values of the hedging point h, the average running cost $C(h)$ can be written in terms of the steady state distribution of the Markov chain. Therefore, the optimal hedging point can be obtained by varying different values of h. Let us first define the following parameters for the manufacturing systems as follows (see Ching and Zhou [6]):

(i) q, the number of machines,
(ii) $1/\sigma_{j1}$, the mean up time of the machine j, $j = 1, \ldots, q$,
(iii) $1/\sigma_{j2}$, the mean repair time for the machine j, $j = 1, \ldots, q$,
(iv) $1/\lambda_j$, the mean processing time for one unit of product on machine j, $j = 1, \ldots, q$,
(v) $1/\mu$, the mean interarrival time of demand,
(vi) h, the hedging point, and
(vii) g, the maximum allowable backlog.

For each machine j, $j = 1, \ldots, q$, let Q_j be the generator matrix of the machine states and Λ_j be the corresponding production rate matrix. Here

$$Q_j = \begin{pmatrix} \sigma_{j1} & -\sigma_{j2} \\ -\sigma_{j1} & \sigma_{j2} \end{pmatrix} \qquad \text{and} \qquad \Lambda_j = \begin{pmatrix} \lambda_j & 0 \\ 0 & 0 \end{pmatrix}$$

(cf. (1)). Each machine has two states—either "up" or "down." Since there are q machines, there are 2^q states for the system of machine. We denote the set of machine states by Ω. The superposition of the q machines forms an MMPP and is characterized by the following $2^q \times 2^q$ generator matrix:

$$Q = (Q_1 \otimes I_2 \otimes \cdots \otimes I_2) + (I_2 \otimes Q_2 \otimes I_2 \otimes \cdots \otimes I_2) + \cdots + (I_2 \otimes \cdots \otimes I_2 \otimes Q_q)$$

(cf. (2)). The corresponding production rate matrix is given by

$$\Lambda = (\Lambda_1 \otimes I_2 \otimes \cdots \otimes I_2) + (I_2 \otimes \Lambda_2 \otimes I_2 \otimes \cdots \otimes I_2) + \cdots + (I_2 \otimes \cdots \otimes I_2 \otimes \Lambda_q)$$

(cf. (3)).

We let $\alpha(t)$ be the state of the system of machines at time t. Therefore, $\alpha(t)$ has 2^q possible states. The inventory level takes integer value in $[-g, h]$ because we allow maximum backlog of g and the hedging point is h. Here negative inventory means backlog. We let $x(t)$ be the inventory level at time t. The machine-inventory process $\{(\alpha(t), x(t)), t \geq 0\}$ forms an irreducible continuous time Markov chain in the state space

$$\{(\alpha, x) \mid \alpha \in \Omega,\ x = -g, \ldots, 0, \ldots, h\}.$$

Each time it visits a state the process stays there for a random period of time that has an exponential distribution and is independent of the past behavior of the process. If we order the state spaces of the machine-inventory process lexicographically, we get the following $(h+g+1)2^q \times (h+g+1)2^q$ generator matrix H for the machine-inventory system:

$$H = \begin{pmatrix} Q+\Lambda & -\mu I & & & & & & 0 \\ -\Lambda & Q+\Lambda+\mu I & -\mu I & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & -\Lambda & Q+\Lambda+\mu I & -\mu I & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & -\Lambda & Q+\Lambda+\mu I & -\mu I & \\ 0 & & & & & -\Lambda & Q+\mu I \end{pmatrix},$$

where I is the $2^m \times 2^m$ identity matrix. Clearly, the matrix H has the same tensor block structure as that of the generator matrix A in (4). In fact, H is a particular case

of A with $s = 1$, $\lambda = 0$, and $m = h + g - 1$. Therefore, the techniques and algorithms developed in the previous sections can be used to obtain the steady state distribution of the process efficiently. Numerical results are given in section 7 to illustrate the fast convergence.

Important quantities such as the average running cost of the machine-inventory system can be written in terms of its steady state distribution. Let

$$p(\alpha, x) = \lim_{t \to \infty} \text{Prob}\{\alpha(t) = \alpha, x(t) = x\}$$

be the steady state probability distribution, and let

$$p_j = \sum_{k \in \Omega} p(k, j), \quad j = -g, -(g-1), \ldots, 0, \ldots, h$$

be the steady state distribution of the inventory level of the system. The average running cost for the machine-inventory system is then given by

$$C(h) = c_I \sum_{j=1}^{h} j p_j - c_B \sum_{j=-g}^{-1} j p_j + c_P \mu p_{-g}, \qquad 0 \le h \le b, \tag{23}$$

where c_I is the inventory cost per unit of product, c_B is the backlog cost per unit of product, c_P is the penalty cost for rejecting an arrival demand, and b is the maximum inventory capacity; see Ching and Zhou [6]. Hence once p_j are given, we can easily find h^* which minimizes the average running cost function $C(h)$ by evaluating $C(h)$ for all $0 \le h \le b$.

We remark that our method can be generalized to handle the case in which each machine has the Erlangian distribution of l phases. Suppose the mean times of repair for machine $j, j = 1, \ldots, q$, are the same in each phase and are equal to $1/\sigma_{j2}$. In this case, the generator matrix for the machine-inventory system can be obtained by replacing the generator matrix of the machine state and its corresponding production rate matrix by $\bar{Q}_j$ and $\bar{\Lambda}_j$, respectively, where

$$\bar{Q}_j = \begin{pmatrix} \sigma_{j1} & & & & -\sigma_{j2} \\ -\sigma_{j1} & \sigma_{j2} & & & \\ & -\sigma_{j2} & \sigma_{j2} & & \\ & & \ddots & \ddots & \\ 0 & & & -\sigma_{j2} & \sigma_{j2} \end{pmatrix} \quad \text{and} \quad \bar{\Lambda}_j = \begin{pmatrix} \lambda_i & & & & 0 \\ & 0 & & & \\ & & 0 & & \\ & & & \ddots & \\ 0 & & & & 0 \end{pmatrix}.$$

Hence we see that the techniques and algorithms developed previously can be applied to this case too.

7. Numerical results. In this section, we illustrate the fast convergence rate of our method by examples in queueing systems and manufacturing systems. The conjugate gradient squared (CGS) method (see Sonneveld [18]) is used to solve the preconditioned system (6). The method does not require the transpose of the iteration matrix GC^{-1}. Using the folding algorithm, one can obtain the steady state probability vector with a residual error of order 10^{-13} to 10^{-16}; see Ye and Li [22]. In order to compare our method with the folding algorithm, the stopping criterion for the CGS and BGS methods is set to be

$$||A\mathbf{p}_k||_2 \le 10^{-12},$$

where $\mathbf{p}_k$ is the computed steady state probability distribution at the kth iteration and

$$||(y_1, y_2, \ldots, y_n)^t||_2 \equiv \sqrt{\sum_{i=1}^{n} y_i^2}.$$

In all our numerical examples, the residual errors lie between 10^{-13} to 10^{-16}, which is comparable to the folding algorithm; see Ye and Li [22]. The initial guess for both methods is the vector of all ones normalized such that its l_2-norm is equal to 1. All the computations were done on an HP 712/80 workstation with MATLAB.

Let us first give the numerical results for the queueing networks. We compare the numerical results of the CGS, preconditioned CGS, and BGS methods for the number of overflow queues $q = 1, 2, 3, 4$ and the number of servers $s = 2$. The MMPP parameters are arbitrarily chosen to be $\sigma_{j1} = 2/3, \sigma_{j2} = 1/3, j = 1, \ldots, q$. The other queueing parameters are given by $\mu = 2$, $\lambda = 1$, and $\lambda_j = 1/q, j = 1, \ldots, q$. We recall that the size of the matrix is $(s+m+1)2^q \times (s+m+1)2^q$. The number of iterations required for convergence is given in Table 1. The symbols I, C, and BGS represent the methods used, namely, CGS without preconditioner, CGS with our preconditioner C in (16), and the block Gauss–Seidel method, respectively. Numbers of iterations greater than 2000 are signified by "$**$."

TABLE 1
Number of iterations for convergence.

$s=2$	$q=1$			$q=2$			$q=3$			$q=4$		
m	I	C	BGS	I	C	BGS	I	C	BGS	I	C	BGS
16	36	7	130	36	9	112	38	12	107	40	13	110
32	155	8	171	154	9	143	158	12	145	161	13	137
64	$**$	7	242	$**$	9	207	$**$	12	213	$**$	13	199
128	$**$	8	366	$**$	10	325	$**$	12	340	$**$	14	317
256	$**$	8	601	$**$	10	549	$**$	12	582	$**$	14	530
512	$**$	8	1051	$**$	10	988	$**$	12	1046	$**$	14	958
1024	$**$	8	$**$	$**$	10	$**$	$**$	12	$**$	$**$	14	$**$

We see that the numbers are roughly constant independent of m for the CGS method with our preconditioner C. For the BGS method, the convergence rate is approximately linear in m. Recall from section 5 that the costs per iteration of the CGS method with preconditioning and of the BGS method are, respectively, $O(2^q(s+m+1)\log_2(s+m+1))$ and $O(2^{2q}(s+m+1))$ operations. We conclude that the total cost of obtaining the steady state probability distribution vector for the CGS method with preconditioning is approximately $O(2^q(s+m+1)\log_2(s+m+1))$ operations while for the BGS method it is approximately $O(2^{2q}m(s+m+1))$ operations.

We next compare the flop counts between our PCG method and the folding algorithm for the single server case ($s = 1$). For simplicity, we set $(s+m+1) = 2^q$ and we consider $q = 1, 2, \ldots, 7$. Our PCG method converges within 25 iterations for all the numerical examples tested. We recall that the number of operations in each iteration of PCG is

$$2\{6(s+m+1)2^q \log_2(s+m+1) + 4(s+m+1)q2^q\}.$$

Therefore, the total number of operations is at most

$$50\{6(s+m+1)2^q \log_2(s+m+1) + 4(s+m+1)q2^q\}.$$

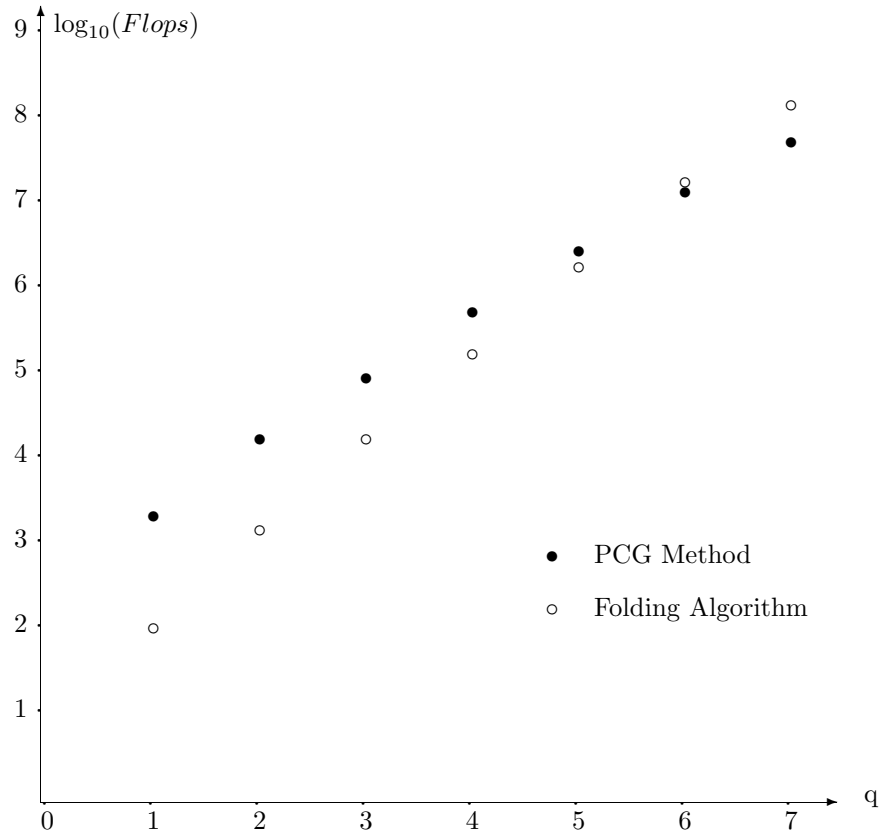


FIG. 1. *Computational flops of the PCG method and the folding algorithm for the single server case.*

The minimum cost of the folding algorithm is given by

$$\frac{26}{3}2^{3q}\log_2(s+m+1)+3(s+m+1)2^{2q};$$

see Ye and Li [22]. In Figure 1, we depict the computational costs of our PCG method and the folding algorithm for different values of q. We see that the computational cost of our PCG method increases at a slower rate than that of the folding algorithm. The crossover point is at $q = 6$.

Next we test our algorithm for the failure-prone manufacturing systems. We assume that all q machines are identical, and in each month (four weeks) each machine breaks down once on average. The mean repairing time for a machine is one week. Therefore, we have $\sigma_{j1} = 1/3, \sigma_{j2} = 1, j = 1, \ldots, q$. The mean time for the arrival of demand is 1/5 week and the mean time for the machine system to produce one unit of product is one day; therefore, we have $\mu = 5$ and $\lambda_j = 7/q, j = 1, \ldots, q$.

In Table 2, we give the number of iterations required for convergence for all three methods. As in the queueing systems case, we see also that the numbers are roughly constant independent of $(g + h)$ for the CGS method with our preconditioner C. For the BGS method, the convergence rate is again approximately linear in $(g + h)$.

Finally, we consider examples of finding the optimal hedging point h^*. We keep the values of the machine parameters the same as in the manufacturing system example above, except that we set $q = 4$ and $g = 50$. Moreover, the inventory cost c_I

TABLE 2
Number of iterations for convergence.

	$q=1$			$q=2$			$q=3$			$q=4$		
$g+h$	I	C	BGS	I	C	BGS	I	C	BGS	I	C	BGS
16	52	6	565	54	7	603	60	8	601	63	9	685
32	173	6	1491	177	8	1682	231	8	1443	180	10	1904
64	**	6	**	**	8	**	**	9	**	**	10	**
128	**	8	**	**	8	**	**	9	**	**	10	**
256	**	8	**	**	9	**	**	9	**	**	10	**
512	**	8	**	**	9	**	**	9	**	**	11	**
1024	**	8	**	**	9	**	**	9	**	**	11	**

TABLE 3
The optimal $(h^*, C(h^*))$ for different λ_i and μ.

	$\mu=1$	$\mu=2$	$\mu=3$
$\lambda_i=1$	(3,181)	(10,533)	(200,14549)
$\lambda_i=1.5$	(2,128)	(5,270)	(11,576)

and backlog cost c_B per unit of product are 50 and 2000, respectively; the maximum inventory capacity b is 200; and the penalty cost c_P for rejecting a demand is 20000 (see (23)). In Table 3, we give the optimal pair of values $(h^*, C(h^*))$, the optimal hedging point h^*, and its corresponding average running cost per week $C(h^*)$ for different values of λ_i and μ.

8. Concluding remarks. In this paper, we proposed a fast algorithm for solving the steady state probability distribution for queueing systems with MMPP inputs. The MMPP is commonly used in modeling the inputs of many physical systems; see Heffes and Lucantoni [8] and Meier-Hellstern [13], for instance. Here we related the MMPP to the production process of unreliable manufacturing systems under the hedging point production control. Our algorithm derived for the queueing systems can be applied to obtain the optimal hedging point. Numerical examples were reported to illustrate the fast convergence rate of our algorithm.

For the manufacturing systems, there are two possible generalizations of the model. The maximum allowable backlog g and the number of machines q (with an associated cost) can be considered as decision variables for the optimization problem. We can also consider the machine failure rate σ_{j1} to be dependent on the production rate λ_j. Note that in this case it has been shown that the optimal policy is still of hedging point type if σ_{j1} is a linear function of the production rate λ_j in the one-machine case; see Hu, Vakili, and Yu [12]. It would be interesting to extend our method to these two cases.

Acknowledgment. The authors would like to thank the referees for their constructive comments and helpful suggestions in revising the paper.

REFERENCES

[1] R. AKELLA AND P. KUMAR, *Optimal control of production rate in a failure prone manufacturing system*, IEEE Trans. Automat. Control, 31 (1986), pp. 116–126.
[2] O. AXELSSON AND V. BARKER, *Finite Element Solution of Boundary Value Problems, Theory and Computation*, Academic Press, New York, 1984.
[3] T. BIELECKI AND P. KUMAR, *Optimality of zero-inventory policies for unreliable manufacturing systems*, Oper. Res., 36 (1988), pp. 532–541.

[4] R. Chan, *Iterative methods for overflow queueing models* I, Numer. Math., 51 (1987), pp. 143–180.
[5] R. Chan, *Circulant preconditioners for Hermitian Toeplitz systems*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 542–550.
[6] W. Ching and X. Zhou, *Matrix methods for production planning in failure prone manufacturing systems*, Lecture Notes in Control and Inform. Sci. 214, Springer-Verlag, London, 1996, pp. 2–30.
[7] P. Davis, *Circulant Matrices*, John Wiley, New York, 1979.
[8] H. Heffes and D. Lucantoni, *A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance*, IEEE J. Select. Areas Commun., SAC-4 (1986), pp. 856–868.
[9] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, London, 1985.
[10] J. Hu and D. Xiang, *The queueing equivalence to optimal control of a manufacturing system with failures*, IEEE Trans. Automat. Control, 38 (1993), pp. 499–502.
[11] J. Hu, *Production rate control for failure prone production systems with no backlog permitted*, IEEE Trans. Automat. Control, 40 (1995), pp. 291–295.
[12] J. Hu, P. Vakili, and G. Yu, *Optimality of hedging point policies in the production control of failure prone manufacturing systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 1875–1880.
[13] K. Meier-Hellstern, *The analysis of a queue arising in overflow models*, IEEE Trans. Commun., 37 (1989), pp. 367–372.
[14] Y. Monden, *Toyota Production System*, Industrial Eng. Manufacturing Press, Atlanta, GA, 1983.
[15] C. Samaratunga, S. Sethi, and X. Zhou, *Computational evaluation of hierarchical production control policies for stochastic manufacturing systems*, Oper. Res., 45 (1997), to appear.
[16] S. Sethi, H. Yan, Q. Zhang, and X. Zhou, *Feedback production planning in a stochastic two-machine flowshop: Asymptotic analysis and computational results*, Internat. J. Production Econ., 30–31 (1993), pp. 79–93.
[17] S. Sethi, Q. Zhang, and X. Zhou, *Hierarchical controls in stochastic manufacturing systems with machines in tandem*, Stochastics Stochastics Rep., 41 (1992), pp. 89–118.
[18] P. Sonneveld, *CGS, a fast Lanczos-type solver for non-symmetric linear systems*, SIAM J. Sci. Comput., 10 (1989), pp. 36–52.
[19] H. Van der Vorst, *Preconditioning by Incomplete Decomposition*, Ph.D. thesis, Rijksuniversiteit te Utrecht, The Netherlands, 1982.
[20] R. Varga, *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1963.
[21] H. Yan, X. Zhou, and G. Yin, *Finding optimal number of kanbans in a manufacturing system via stochastic approximation and perturbation analysis*, in Lecture Notes in Control and Information Sciences, G. Cohen and J.-P. Quadrat, eds., Springer-Verlag, London, 1994, pp. 572–578.
[22] J. Ye and S. Li, *Analysis of multi-media traffic queues with finite buffer and overload control, Part* I: *Algorithm*, Proc. IEEE INFOCOM '91, (1991), pp. 1464–1474.
[23] H. Young, B. Byung, and K. Chong, *Performance analysis of leaky-bucket bandwidth enforcement strategy for bursty traffics in an ATM network*, Comput. Net. ISDN Syst., 25 (1992), pp. 295–303.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, pp. 482–498, April 1997

016

QUASI LUMPABILITY, LOWER-BOUNDING COUPLING MATRICES, AND NEARLY COMPLETELY DECOMPOSABLE MARKOV CHAINS*

TUĞRUL DAYAR† AND WILLIAM J. STEWART‡

Abstract. In this paper, it is shown that nearly completely decomposable (NCD) Markov chains are quasi-lumpable. The state space partition is the natural one, and the technique may be used to compute lower and upper bounds on the stationary probability of each NCD block. In doing so, a lower-bounding nonnegative coupling matrix is employed. The nature of the stationary probability bounds is closely related to the structure of this lower-bounding matrix. Irreducible lower-bounding matrices give tighter bounds compared with bounds obtained using reducible lower-bounding matrices. It is also noticed that the quasi-lumped chain of an NCD Markov chain is an ill-conditioned matrix and the bounds obtained generally will not be tight. However, under some circumstances, it is possible to compute the stationary probabilities of some NCD blocks exactly.

Key words. Markov chains, quasi lumpability, decomposability, stationary probability, aggregation–disaggregation schemes

AMS subject classifications. 60J10, 60J27, 65U05, 65F05, 65F10, 65F30

PII. S0895479895294277

1. Introduction.

Markovian modeling and analysis are extensively used in many disciplines in evaluating the performance of existing systems and in analyzing and designing systems to be developed. The long-run behavior of Markovian systems is revealed through the solution of the problem

$$\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}, \ \|\boldsymbol{\pi}\|_1 = 1, \tag{1.1}$$

where $\mathbf{P}$ is the one-step stochastic transition probability matrix (i.e., discrete-time Markov chain—DTMC) and $\boldsymbol{\pi}$ is the unknown stationary probability distribution of the system under consideration. By definition, rows of $\mathbf{P}$ and elements of $\boldsymbol{\pi}$ both sum up to 1.

In what follows, boldface capital letters denote matrices, boldface lowercase letters denote column vectors, italic lowercase and uppercase letters denote scalars, and calligraphic letters denote sets. $\boldsymbol{e}$ represents a column vector of all ones and $\mathbf{o}$ represents a row or column vector of all zeros depending on the context. The convention of representing probability distributions by row vectors is adopted.

Solving (1.1) is crucial in computing performance measures for Markovian systems. For queueing systems, these measures may be the average number of customers, the mean waiting time, or the blocking probability for a specific queue. In communication systems, they may be the total packet loss rate, the probability of an empty system, or any other relevant measure. In any case, these measures may be computed exactly if $\boldsymbol{\pi}$ is available.

* Received by the editors November 1, 1995; accepted for publication (in revised form) by D. P. O'Leary June 11, 1996. This work was initiated while T. Dayar was in the Department of Computer Science at North Carolina State University. His work is currently supported by Scientific and Technical Research Council of Turkey (TÜBİTAK) grant EEEAG-161.

http://www.siam.org/journals/simax/18-2/29427.html

† Department of Computer Engineering and Information Science, Bilkent University, 06533 Bilkent, Ankara, Turkey (tugrul@bilkent.edu.tr).

‡ Department of Computer Science, North Carolina State University, Raleigh, NC 27695-8206 (billy@markov.csc.ncsu.edu).

NCD Markov chains [3], [10], [16] are irreducible stochastic matrices that can be ordered so that the matrix of transition probabilities has a block structure in which the nonzero elements of the off-diagonal blocks are small compared with those of the diagonal blocks. Such matrices often arise in queueing network analysis, large-scale economic modeling, and computer systems performance evaluation, and they can be represented in the form

$$\mathbf{P}_{n\times n} = \begin{array}{c} \begin{array}{cccc} n_1 & n_2 & \cdots & n_N \end{array} \\ \begin{pmatrix} \mathbf{P}_{1,1} & \mathbf{P}_{1,2} & \cdots & \mathbf{P}_{1,N} \\ \mathbf{P}_{2,1} & \mathbf{P}_{2,2} & \cdots & \mathbf{P}_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}_{N,1} & \mathbf{P}_{N,2} & \cdots & \mathbf{P}_{N,N} \end{pmatrix} \end{array} \begin{array}{c} n_1 \\ n_2 \\ \vdots \\ n_N \end{array} . \tag{1.2}$$

The subblocks $\mathbf{P}_{i,i}$ are square, of order n_i, with $n = \sum_{i=1}^{N} n_i$. Let $\boldsymbol{\pi}$ be partitioned conformally with $\mathbf{P}$ such that $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_N)$. Each $\boldsymbol{\pi}_i$, $i = 1, 2, \ldots, N$ is a row vector having n_i elements. Let $\mathbf{P} = \operatorname{diag}(\mathbf{P}_{1,1}, \mathbf{P}_{2,2}, \ldots, \mathbf{P}_{N,N}) + \mathbf{E}$. The quantity $\|\mathbf{E}\|_\infty$ is referred to as the degree of coupling, and it is taken to be a measure of the decomposability of the matrix (see [6]). If it were zero, then $\mathbf{P}$ would be reducible.

Consider the following questions. Is it possible to obtain lower and upper bounds on the stationary probability of being in each NCD block of an NCD Markov chain in an inexpensive way? Furthermore, if the answer to the preceding question is yes, can one improve these bounds by exploiting the structure and symmetries of the chain? The motivation behind seeking answers to such questions is that in many cases performance measures of interest of systems undergoing analysis depend on the probability of being in certain groups of states. That is, probabilities need to be computed at a coarser level; each and every stationary probability is not needed. If the problem at hand is one in which the stationary probabilities of interest are those of the coupling matrix [10] corresponding to the underlying NCD Markov chain, then the technique discussed in this paper may be used to obtain answers to the above questions. Whereas if all stationary probabilities of the NCD Markov chain are to be computed, iterative aggregation–disaggregation (IAD) should be the method of choice (see [8], [2], [12], [15], [14], [16]).

In the sections to come, it is shown that NCD Markov chains are quasi-lumpable. The state space partition coincides with the NCD block partition, and the technique may be used to compute lower and upper bounds on the probability of being in each NCD block. The procedure amounts to solving linear systems of order equal to the number of NCD blocks in the chain. Thereafter, quasi lumpability is related to the polyhedra theory of Courtois and Semal for stochastic matrices [4], and it is shown that under certain circumstances the quasi-lumped chain (as defined in [5]) is a lower-bounding matrix for the coupling matrix of the NCD chain. Additionally, another substochastic matrix guaranteed to be a lower-bounding coupling matrix is given. Following this, the effects of the nonzero structure of a lower-bounding nonnegative coupling matrix on the bounds of the stationary probability of each NCD block is investigated; the results are based on the nonzero structure of a lower-bounding substochastic matrix in general, and, therefore, they may also be used in forecasting the quality of lower and upper bounds on the stationary distribution of Markov chains when Courtois and Semal's theory is at work.

The next section provides the definitions of lumpability (see [7, section 6.3]) and quasi lumpability (see [5]), and section 3 shows how quasi lumpability applies to NCD

Markov chains. The effects of quasi lumpability on the 8×8 Courtois matrix are illustrated in section 4. The relation between the quasi-lumped chain and the coupling matrix of an NCD Markov chain is investigated in section 5. Section 6 provides information enabling one to forecast the nature of the bounds on the stationary probability of each NCD block; the idea is communicated through an illustrative example. The last section summarizes the results.

2. Lumpability vs. quasi lumpability. Lumpability is a property of some Markov chains which, if conditions are met, may be used to reduce a large state space to a smaller one. The idea is to find a partition of the original state space such that, when the states in each partition are combined to form a single state, the resulting Markov chain described by the combined states has equivalent behavior to the original chain, only at a coarser level of detail. Given that the conditions for lumpability are satisfied, it is mostly useful in systems which require the computation of performance measures dependent on the coarser analysis specified by the lumped chain (see [7, p. 123]).

DEFINITION 2.1. *A DTMC is said to be* lumpable *with respect to a given state space partition* $\mathcal{S} = \bigcup_i \mathcal{S}_i$ *with* $\mathcal{S}_i \bigcap \mathcal{S}_j = \emptyset \ \forall i \neq j$ *if its transition probability matrix* $\mathbf{P}$ *satisfies the lumpability condition*

$$\forall \mathcal{S}_i, \mathcal{S}_j \subset \mathcal{S} \ \forall s \in \mathcal{S}_i : \sum_{s' \in \mathcal{S}_j} p_{s,s'} = k_{i,j} \qquad \forall i, j, \tag{2.1}$$

where $k_{i,j}$ *is a constant value that depends only on* i *and* j *and* $p_{s,s'}$ *is the one-step transition probability of going from state* s *to state* s'. *The* lumped chain $\mathbf{K}$ *has* $k_{i,j}$ *as its* i, j*th entry. A similar definition applies to a continuous-time Markov chain (CTMC), where the probability matrix* $\mathbf{P}$ *is substituted with the infinitesimal generator* $\mathbf{Q}$.

To put it in another way, the lumpability condition requires the transition probability from each state in a given partition to another partition to be the same. For a given state, the probability of making a transition to a partition is the sum of the transition probabilities from the given state to each state in that partition. At this point we should stress that not all Markov chains are lumpable. In fact, only a small percentage of Markov chains arising in real-life applications is expected to be lumpable. However, in section 3 it is shown that NCD Markov chains are quasi-lumpable, that is, almost lumpable [5]. The following informative example demonstrates the concept of lumpability.

Example 2.2. Let

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0.2 & 0.3 & 0.4 & 0.1 \\ 0.3 & 0.1 & 0.4 & 0.2 \\ 0.5 & 0.1 & 0.1 & 0.3 \\ 0.5 & 0.3 & 0.2 & 0 \end{pmatrix} \end{array}.$$

We take partition $\mathcal{S} = \{1, 3\} \bigcup \{2, 4\}$. For this partition the lumpability condition is satisfied with $k_{1,1} = 0.6, k_{1,2} = 0.4, k_{2,1} = 0.7, k_{2,2} = 0.3$, where $\mathcal{S}_1 = \{1, 3\}, \mathcal{S}_2 = \{2, 4\}$. The lumped chain is given by

$$\mathbf{K} = \begin{array}{c} \\ \mathcal{S}_1 \\ \mathcal{S}_2 \end{array} \begin{array}{c} \begin{array}{cc} \mathcal{S}_1 & \mathcal{S}_2 \end{array} \\ \begin{pmatrix} 0.6 & 0.4 \\ 0.7 & 0.3 \end{pmatrix} \end{array}.$$

DEFINITION 2.3. *A DTMC is said to be* ϵ quasi-lumpable *with respect to a given state space partition* $\mathcal{S} = \bigcup_i \mathcal{S}_i$ *with* $\mathcal{S}_i \bigcap \mathcal{S}_j = \emptyset \ \forall i \neq j$ *if its transition probability matrix* $\mathbf{P}$ *can be written as* $\mathbf{P} = \mathbf{P}^- + \mathbf{P}^\epsilon$. *Here* $\mathbf{P}^-$ *is a (componentwise) lower bound for* $\mathbf{P}$ *that satisfies the lumpability condition*

$$\forall \mathcal{S}_i, \mathcal{S}_j \subset \mathcal{S} \ \forall s \in \mathcal{S}_i : \ \sum_{s' \in \mathcal{S}_j} p^-_{s,s'} = k_{i,j} \qquad \forall i \neq j \tag{2.2}$$

under the following constraints. No element in $\mathbf{P}^\epsilon$ *is greater than* ϵ *(a small number);* $\|\mathbf{P}^\epsilon\|_\infty$ *assumes the minimum value among all possible alternatives (since* $\mathbf{P}^-$ *and* $\mathbf{P}^\epsilon$ *may not be unique);* $k_{i,j}$ *is a constant value that depends only on* i *and* j*; and* $p^-_{s,s'}$ *is the one-step transition probability of going from state* s *to state* s' *in the matrix* $\mathbf{P}^-$ *(see* [5, *p.* 224]*). The computation of the quasi-lumped chain is discussed in the next section. A similar definition applies to a CTMC as in Definition* 2.1.

The concept of ϵ *quasi lumpability* is illustrated in the following 6×6 example.

Example 2.4. Let

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \left(\begin{array}{cccc|cc} 0.2 & 0.28 & 0.1 & 0.21 & 0.11 & 0.1 \\ 0.29 & 0.1 & 0.2 & 0.05 & 0.31 & 0.05 \\ 0.15 & 0.2 & 0.24 & 0.12 & 0.2 & 0.09 \\ \hline 0.27 & 0.18 & 0.22 & 0.18 & 0.01 & 0.14 \\ 0.18 & 0.2 & 0.3 & 0.31 & 0.01 & 0 \\ 0 & 0.25 & 0.43 & 0.07 & 0.08 & 0.17 \end{array} \right) \end{array}.$$

$\mathbf{P}^-$ and $\mathbf{P}^\epsilon$ given by

$$\mathbf{P}^- = \left(\begin{array}{cccc|cc} 0.2 & 0.28 & 0.1 & 0.2 & 0.11 & 0.1 \\ 0.29 & 0.1 & 0.2 & 0.05 & 0.31 & 0.05 \\ 0.15 & 0.2 & 0.24 & 0.12 & 0.2 & 0.09 \\ \hline 0.27 & 0.18 & 0.22 & 0.18 & 0.01 & 0.14 \\ 0.18 & 0.2 & 0.29 & 0.31 & 0.01 & 0 \\ 0 & 0.24 & 0.43 & 0.07 & 0.08 & 0.17 \end{array} \right),$$

$$\mathbf{P}^\epsilon = \left(\begin{array}{ccc|cc} 0 & 0 & 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.01 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 & 0 & 0 \end{array} \right)$$

with $\epsilon = 0.01$ and state space partition $\mathcal{S} = \{1,2,3\} \bigcup \{4,5,6\}$ satisfy the quasi-lumpability condition in (2.2). This time $\mathcal{S}_1 = \{1,2,3\}, \mathcal{S}_2 = \{4,5,6\}$, and $k_{1,2} = 0.41, k_{2,1} = 0.67$. Observe that for $\epsilon = 0.01$ the given ($\mathbf{P}^-$, $\mathbf{P}^\epsilon$) pair is not the only one that satisfies the quasi-lumpability condition. For instance, the following pair also satisfies (2.2):

$$\mathbf{P}^- = \left(\begin{array}{cccc|cc} 0.2 & 0.28 & 0.1 & 0.21 & 0.1 & 0.1 \\ 0.29 & 0.1 & 0.2 & 0.05 & 0.31 & 0.05 \\ 0.15 & 0.2 & 0.24 & 0.12 & 0.2 & 0.09 \\ \hline 0.27 & 0.18 & 0.22 & 0.18 & 0.01 & 0.14 \\ 0.17 & 0.2 & 0.3 & 0.31 & 0.01 & 0 \\ 0 & 0.25 & 0.42 & 0.07 & 0.08 & 0.17 \end{array} \right),$$

$$\mathbf{P}^{\epsilon} = \left(\begin{array}{ccc|ccc} 0 & 0 & 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0.01 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.01 & 0 & 0 & 0 \end{array}\right).$$

The next section provides a proof by construction for the ϵ quasi lumpability of NCD Markov chains.

3. Construction.

1. For an NCD Markov chain, let the state space be partitioned as

$$\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_N\},$$

where $\mathcal{S}_i$ is the set of states forming the ith block and $\#(\mathcal{S}_i) = n_i$ with $n = \sum_{i=1}^{N} n_i$.

Form the matrix

$$\mathbf{P}^{-} = \begin{array}{c} \begin{array}{cccc} n_1 & n_2 & \cdots & n_N \end{array} \\ \left(\begin{array}{cccc} \mathbf{P}_{1,1} & \mathbf{P}^{-}_{1,2} & \cdots & \mathbf{P}^{-}_{1,N} \\ \mathbf{P}^{-}_{2,1} & \mathbf{P}_{2,2} & \cdots & \mathbf{P}^{-}_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}^{-}_{N,1} & \mathbf{P}^{-}_{N,2} & \cdots & \mathbf{P}_{N,N} \end{array}\right) \end{array} \begin{array}{c} n_1 \\ n_2 \\ \vdots \\ n_N \end{array}, \tag{3.1}$$

where

$$\mathbf{P}^{-}_{i,j} = \begin{cases} \mathbf{P}_{i,j} & \text{if } \mathbf{P}_{i,j}\boldsymbol{e} = k_{i,j}\boldsymbol{e} \\ \mathbf{P}_{i,j} - \mathbf{P}^{\epsilon}_{i,j} & \text{otherwise} \end{cases} \qquad \forall i \neq j. \tag{3.2}$$

Diagonal blocks of $\mathbf{P}^{-}$ are the same as those of $\mathbf{P}$. When $\mathbf{P}_{i,j}\boldsymbol{e} \neq k_{i,j}\boldsymbol{e}$, $\mathbf{P}^{\epsilon}_{i,j}$ is chosen so that $(\mathbf{P}_{i,j} - \mathbf{P}^{\epsilon}_{i,j})\boldsymbol{e} = k_{i,j}\boldsymbol{e}$. Here, $k_{i,j} = \min(\mathbf{P}_{i,j}\boldsymbol{e})$ (i.e., the minimum-valued element of the vector $\mathbf{P}_{i,j}\boldsymbol{e}$). As pointed out in Example 2.4, $\mathbf{P}^{\epsilon}$ may not be unique, and the discussion on how to choose among the alternatives available is left to after the construction. Furthermore, $\mathbf{P}^{\epsilon}$ has nonzero blocks (in which there is at least one nonzero element) in locations corresponding to the nonzero blocks of $\mathbf{P}$ which do not have equal row sums. On the other hand, the number of zero blocks in $\mathbf{P}^{-}$ may be more than the number of zero blocks in $\mathbf{P}$. In other words, there may be nonzero blocks in $\mathbf{P}$ for which $k_{i,j} = 0$, implying $\mathbf{P}^{-}_{i,j} = \mathbf{0}$. Note that, if $\mathbf{P}^{\epsilon}$ is the null matrix, then $\mathbf{P}$ will be exactly lumpable, and the remaining steps in the construction should be skipped.

2. Once $\mathbf{P}$ is written as the sum of $\mathbf{P}^{-}$ and $\mathbf{P}^{\epsilon}$, form yet another matrix

$$\mathbf{P}^{s} = \begin{pmatrix} \mathbf{P}^{-} & \boldsymbol{y} \\ \boldsymbol{x}^{T} & 0 \end{pmatrix}, \tag{3.3}$$

where

$$\boldsymbol{y} = \begin{pmatrix} \bar{\boldsymbol{y}}_1 \\ \bar{\boldsymbol{y}}_2 \\ \vdots \\ \bar{\boldsymbol{y}}_N \end{pmatrix} = \mathbf{P}^{\epsilon}\boldsymbol{e} \tag{3.4}$$

and $\bar{\boldsymbol{y}}_i$ has n_i elements. The unknown vector $\boldsymbol{x}$ should be partitioned in the same way. The significance and role of $\boldsymbol{x}$ in the computation of lower and upper bounds for the quasi-lumped chain is discussed in section 4. Recall the definition of an NCD Markov chain in section 1 and observe that $\|\boldsymbol{y}\|_\infty \leq \|\mathbf{E}\|_\infty$ (the degree of coupling of $\mathbf{P}$). Since $\|\mathbf{E}\|_\infty$ is a small number generally less than 0.1, one has ϵ quasi lumpability (see Definition 2.3). The small mass in the off-diagonal blocks, which prevents lumping $\mathbf{P}$ exactly, is accumulated in an extra state.

3. Given that $\mathbf{P}$ is not exactly lumpable (i.e., $\boldsymbol{y} \neq \mathbf{o}$), $\mathbf{P}^s$ will not be lumpable. However, the lumpability condition for the ith row of blocks may be enforced by increasing some elements in $\bar{\boldsymbol{y}}_i$ so as to make each element equal to $\|\bar{\boldsymbol{y}}_i\|_\infty$ and decreasing the corresponding diagonal elements. If it is possible for any diagonal element to become negative, the diagonal of $\mathbf{P}^s$ may be scaled by performing the transformation

$$\alpha\mathbf{P}^s + (1-\alpha)\mathbf{I}, \tag{3.5}$$

where $0 < \alpha < 1$, on $\mathbf{P}^s$ as suggested in [5]. Denote the matrix obtained in the end $\tilde{\mathbf{P}}^s$.

4. $\tilde{\mathbf{P}}^s$ is lumpable, and it may be lumped to form the following *quasi-lumped chain* that corresponds to $\mathbf{P}$:

$$\mathbf{K}^s = \left(\begin{array}{cccc|c} \|\tilde{\mathbf{P}}^s_{1,1}\|_\infty & \|\tilde{\mathbf{P}}^s_{1,2}\|_\infty & \cdots & \|\tilde{\mathbf{P}}^s_{1,N}\|_\infty & \|\bar{\boldsymbol{y}}_1\|_\infty \\ \|\tilde{\mathbf{P}}^s_{2,1}\|_\infty & \|\tilde{\mathbf{P}}^s_{2,2}\|_\infty & \cdots & \|\tilde{\mathbf{P}}^s_{2,N}\|_\infty & \|\bar{\boldsymbol{y}}_2\|_\infty \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \|\tilde{\mathbf{P}}^s_{N,1}\|_\infty & \|\tilde{\mathbf{P}}^s_{N,2}\|_\infty & \cdots & \|\tilde{\mathbf{P}}^s_{N,N}\|_\infty & \|\bar{\boldsymbol{y}}_N\|_\infty \\ \hline \|\bar{\boldsymbol{x}}_1\|_1 & \|\bar{\boldsymbol{x}}_2\|_1 & \cdots & \|\bar{\boldsymbol{x}}_N\|_1 & 0 \end{array}\right). \tag{3.6}$$

5. Bounds on the stationary probability of each NCD block may be obtained using Courtois and Semal's method [4], [13] if the $N \times N$ principal submatrix of $\mathbf{K}^s$ is a lower-bounding coupling matrix for $\mathbf{P}$.

When constructing $\mathbf{P}^\epsilon$, the nonzero elements in blocks should be arranged, if at all possible, so that there is a minimum number of nonzero columns in $\mathbf{P}^\epsilon$. If all columns corresponding to states in $\mathcal{S}_i$ are zero in $\mathbf{P}^\epsilon$, then $\bar{\boldsymbol{x}}_i = \mathbf{o}$, and the stationary probability of the ith block may be determined exactly to working precision. An intuitive explanation for this fact is the following. The transitions in $\mathbf{P}^\epsilon$ are the transitions into and out of the extra state (in $\mathbf{P}^s$). Therefore, if it is not possible to make a transition to state s, say, in the matrix $\mathbf{P}^\epsilon$ (i.e., the column of $\mathbf{P}^\epsilon$ that corresponds to state s is $\mathbf{o}$), then it will not be possible to return to state s from the extra state. This being so, the corresponding element in $\boldsymbol{x}^T$ must be zero. If all states in an NCD block possess this property, then the element in the last row of the quasi-lumped chain $\mathbf{K}^s$ corresponding to that NCD block should be zero. A side-note is that, even though there may be multiple ways in which the nonzero entries of $\mathbf{P}^\epsilon$ can be arranged for fixed ϵ, this does not make a difference when lower and upper bounds on the stationary probability of each NCD block are computed.

The next section illustrates the construction steps on a small example and shows how to compute the corresponding quasi-lumped chain with lower and upper bounds for its stationary vector.

4. An illustrative example. Consider the 8×8 Courtois matrix [3]

$$\mathbf{P} = \left(\begin{array}{ccc|ccc|cc} 0.85 & 0 & 0.149 & 0.0009 & 0 & 0.00005 & 0 & 0.00005 \\ 0.1 & 0.65 & 0.249 & 0 & 0.0009 & 0.00005 & 0 & 0.00005 \\ 0.1 & 0.8 & 0.0996 & 0.0003 & 0 & 0 & 0.0001 & 0 \\ \hline 0 & 0.0004 & 0 & 0.7 & 0.2995 & 0 & 0.0001 & 0 \\ 0.0005 & 0 & 0.0004 & 0.399 & 0.6 & 0.0001 & 0 & 0 \\ \hline 0 & 0.00005 & 0 & 0 & 0.00005 & 0.6 & 0.2499 & 0.15 \\ 0.00003 & 0 & 0.00003 & 0.00004 & 0 & 0.1 & 0.8 & 0.0999 \\ 0 & 0.00005 & 0 & 0 & 0.00005 & 0.1999 & 0.25 & 0.55 \end{array}\right).$$

The degree of coupling for this matrix is 0.001. From the first step of the construction, one obtains

$$\mathbf{P}^{-} = \left(\begin{array}{ccc|ccc|cc} 0.85 & 0 & 0.149 & 0.0003 & 0 & 0.00005 & 0 & 0.00005 \\ 0.1 & 0.65 & 0.249 & 0 & 0.0003 & 0.00005 & 0 & 0.00005 \\ 0.1 & 0.8 & 0.0996 & 0.0003 & 0 & 0 & 0.0001 & 0 \\ \hline 0 & 0.0004 & 0 & 0.7 & 0.2995 & 0 & 0.0001 & 0 \\ 0 & 0 & 0.0004 & 0.399 & 0.6 & 0.0001 & 0 & 0 \\ \hline 0 & 0.00005 & 0 & 0 & 0.00004 & 0.6 & 0.2499 & 0.15 \\ 0.00002 & 0 & 0.00003 & 0.00004 & 0 & 0.1 & 0.8 & 0.0999 \\ 0 & 0.00005 & 0 & 0 & 0.00004 & 0.1999 & 0.25 & 0.55 \end{array}\right),$$

$$\mathbf{P}^{\epsilon} = \left(\begin{array}{ccc|ccc|ccc} 0 & 0 & 0 & 0.0006 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0006 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.0005 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0.00001 & 0 & 0 & 0 \\ 0.00001 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.00001 & 0 & 0 & 0 \end{array}\right).$$

$\mathbf{P} = \mathbf{P}^{-} + \mathbf{P}^{\epsilon}$ (with $\epsilon = 0.0006$), as required, and the second step of the construction gives

$$\mathbf{P}^{s} = \left(\begin{array}{ccc|ccc|ccc} 0.85 & 0 & 0.149 & 0.0003 & 0 & 0.00005 & 0 & 0.00005 & 0.0006 \\ 0.1 & 0.65 & 0.249 & 0 & 0.0003 & 0.00005 & 0 & 0.00005 & 0.0006 \\ 0.1 & 0.8 & 0.0996 & 0.0003 & 0 & 0 & 0.0001 & 0 & 0 \\ \hline 0 & 0.0004 & 0 & 0.7 & 0.2995 & 0 & 0.0001 & 0 & 0 \\ 0 & 0 & 0.0004 & 0.399 & 0.6 & 0.0001 & 0 & 0 & 0.0005 \\ \hline 0 & 0.00005 & 0 & 0 & 0.00004 & 0.6 & 0.2499 & 0.15 & 0.00001 \\ 0.00002 & 0 & 0.00003 & 0.00004 & 0 & 0.1 & 0.8 & 0.0999 & 0.00001 \\ 0 & 0.00005 & 0 & 0 & 0.00004 & 0.1999 & 0.25 & 0.55 & 0.00001 \\ \hline x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & 0 \end{array}\right).$$

Note that there are no transitions to states 2, 3, 6, 7, and 8 in $\mathbf{P}^{\epsilon}$. Hence, x_2, x_3, x_6, x_7, and x_8 in $\mathbf{P}^{s}$ must be zero. Observe that $\mathbf{P}^{s}$ is still not lumpable. For it to be lumpable, the last column should be modified. Following the third step of the construction, diagonal elements $p^{s}_{3,3}$ and $p^{s}_{4,4}$ are adjusted and one obtains

$$\tilde{\mathbf{P}}^{s} = \left(\begin{array}{ccc|ccc|ccc} 0.85 & 0 & 0.149 & 0.0003 & 0 & 0.00005 & 0 & 0.00005 & 0.0006 \\ 0.1 & 0.65 & 0.249 & 0 & 0.0003 & 0.00005 & 0 & 0.00005 & 0.0006 \\ 0.1 & 0.8 & 0.099 & 0.0003 & 0 & 0 & 0.0001 & 0 & 0.0006 \\ \hline 0 & 0.0004 & 0 & 0.6995 & 0.2995 & 0 & 0.0001 & 0 & 0.0005 \\ 0 & 0 & 0.0004 & 0.399 & 0.6 & 0.0001 & 0 & 0 & 0.0005 \\ \hline 0 & 0.00005 & 0 & 0 & 0.00004 & 0.6 & 0.2499 & 0.15 & 0.00001 \\ 0.00002 & 0 & 0.00003 & 0.00004 & 0 & 0.1 & 0.8 & 0.0999 & 0.00001 \\ 0 & 0.00005 & 0 & 0 & 0.00004 & 0.1999 & 0.25 & 0.55 & 0.00001 \\ \hline x_1 & 0 & x_3 & x_4 & x_5 & 0 & 0 & 0 & 0 \end{array}\right).$$

Notice that x_3 in $\tilde{\mathbf{P}}^s$ is different than zero, as opposed to what has been said before. The reason is that $p^s_{3,3}$ has been adjusted, thus making $p^\epsilon_{3,3}$ effectively a nonzero entry of value 0.0006. Therefore, the third column in $\mathbf{P}^\epsilon$ intrinsically has a nonzero entry in the diagonal position, implying a transition from the extra state to state 3. Likewise, $p^s_{4,4}$ has been adjusted, making $p^\epsilon_{4,4}$ equal to 0.0005. However, x_4 is already nonzero and need not be altered. This issue will be revisited at the end of the section. Resuming the construction, the quasi-lumped chain in step four is computed as

$$\mathbf{K}^s = \left(\begin{array}{cccc} 0.999 & 0.0003 & 0.0001 & 0.0006 \\ 0.0004 & 0.999 & 0.0001 & 0.0005 \\ 0.00005 & 0.00004 & 0.9999 & 0.00001 \\ \hline \|\bar{\boldsymbol{x}}_1\|_1 & \|\bar{\boldsymbol{x}}_2\|_1 & 0 & 0 \end{array}\right).$$

As suggested in the fifth step of the construction, lower and upper bounds on the stationary probability of each NCD block may be obtained by successively substituting a one for each (unknown) $\|\bar{\boldsymbol{x}}_i\|_1$ in the last row of $\mathbf{K}^s$ (denote this matrix by $\mathbf{K}^s_i$) and solving the corresponding system

$$\boldsymbol{z}_i \mathbf{K}^s_i = \boldsymbol{z}_i, \quad \sum_{j=1}^{N} z_{i,j} = 1. \tag{4.1}$$

Here, $\boldsymbol{z}_i$ is a probability vector of N elements. If ξ_j is the stationary probability of the jth NCD block, then lower and upper bounds on the stationary probability of block j may be computed from

$$\xi_j^{inf} = \max\left\{\min_i(z_{i,j});\ 1 - \sum_{k\neq j}\max_i(z_{i,k})\right\}, \tag{4.2}$$

$$\xi_j^{sup} = \min\left\{\max_i(z_{i,j});\ 1 - \sum_{k\neq j}\min_i(z_{i,k})\right\} \tag{4.3}$$

(see [4, (3.26), p. 810]).

For the Courtois matrix

$$\|\bar{\boldsymbol{x}}_1\|_1 = 1,\ \|\bar{\boldsymbol{x}}_2\|_1 = 0 \quad \Rightarrow \quad \boldsymbol{z}_1 = [0.36923, 0.13077, 0.50000],$$

$$\|\bar{\boldsymbol{x}}_1\|_1 = 0,\ \|\bar{\boldsymbol{x}}_2\|_1 = 1 \quad \Rightarrow \quad \boldsymbol{z}_2 = [0.16071, 0.33929, 0.50000],$$

$$\begin{aligned} 0.16071 \leq \xi_1 \leq 0.36923, \\ 0.13077 \leq \xi_2 \leq 0.33929, \\ 0.50000 \leq \xi_3 \leq 0.50000 \ \Rightarrow\ \xi_3 = 0.50000, \end{aligned}$$

and $\xi_1 + \xi_2 + \xi_3 = 1$ in five decimal digits of accuracy.

We obtained the stationary probability of each NCD block by solving for the stationary vector of the original 8×8 chain. The probabilities accurate to five decimal digits are

$$\xi_1 = 0.22253, \quad \xi_2 = 0.27747, \quad \xi_3 = 0.50000.$$

The next thing to do is to show how a distribution $\boldsymbol{x}^T$ that gives the stationary probability of each NCD block may be obtained. In fact, the procedure amounts to computing x_1, x_3, x_4, and x_5 values only, for the rest of the elements in $\boldsymbol{x}$ are necessarily zero. Let $\boldsymbol{\pi}$ denote the stationary vector of $\mathbf{P}$ (i.e., $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$, $\|\boldsymbol{\pi}\|_1 = 1$). Then

$$\begin{aligned} x_1 &= (0.0005\pi_5 + 0.00001\pi_7)/t, \\ x_3 &= 0.0006\pi_3/t, \\ x_4 &= (0.0006\pi_1 + 0.0005\pi_4)/t, \\ x_5 &= (0.0006\pi_2 + 0.00001\pi_6 + 0.00001\pi_8)/t, \end{aligned}$$

where $t = 0.0006(\pi_1 + \pi_2 + \pi_3) + 0.0005(\pi_4 + \pi_5) + 0.00001(\pi_6 + \pi_7 + \pi_8)$. The last condition ensures that $\boldsymbol{x}^T$ is a probability vector. As can be seen, the computation of $\boldsymbol{x}$ requires full knowledge of $\boldsymbol{\pi}$ (which, of course, is unknown). For the Courtois matrix, the unknown entries in the last row of $\mathbf{K}^s$ are given by

$$\begin{aligned} \|\bar{\boldsymbol{x}}_1\|_1 &= (0.0006\pi_3 + 0.0005\pi_5 + 0.00001\pi_7)/t, \\ \|\bar{\boldsymbol{x}}_2\|_1 &= (0.0006(\pi_1 + \pi_2) + 0.0005\pi_4 + 0.00001(\pi_6 + \pi_8))/t. \end{aligned}$$

Using $\boldsymbol{\pi}$, one computes $\|\bar{\boldsymbol{x}}_1\|_1 = 0.31213$, $\|\bar{\boldsymbol{x}}_2\|_1 = 0.68787$ in five decimal digits of accuracy as the combination that gives $\boldsymbol{\xi}$.

The next section relates the quasi-lumped chain to the coupling matrix of the original NCD Markov chain.

5. Quasi-lumped chain and the coupling matrix. Let $\mathbf{C}^s$ denote the $N \times N$ principal submatrix of the quasi-lumped chain $\mathbf{K}^s$. For the Courtois matrix,

$$\mathbf{C}^s = \begin{pmatrix} 0.99900 & 0.00030 & 0.00010 \\ 0.00040 & 0.99900 & 0.00010 \\ 0.00005 & 0.00004 & 0.99990 \end{pmatrix}.$$

On the other hand, the entries of the coupling matrix of an NCD Markov chain are given by [11]

$$c_{i,j} = \frac{\boldsymbol{\pi}_i}{\|\boldsymbol{\pi}_i\|_1}\mathbf{P}_{i,j}\boldsymbol{e} \qquad \forall i, j.$$

For the same example, the coupling matrix in five decimal digits of accuracy is then

$$\mathbf{C} = \begin{pmatrix} 0.99911 & 0.00079 & 0.00010 \\ 0.00061 & 0.99929 & 0.00010 \\ 0.00006 & 0.00004 & 0.99990 \end{pmatrix}.$$

In this example, $\mathbf{C}^s$ is a lower bound for the exact coupling matrix $\mathbf{C}$. That is, $\mathbf{C}^s \leq \mathbf{C}$. Is this always true? Before answering this question, two lemmas should be stated. In the following, $\boldsymbol{u} \ll \boldsymbol{v}$ means each element of $\boldsymbol{u}$ is considerably smaller than the corresponding element of $\boldsymbol{v}$. The symbol $\ll$ may also be used between two scalars (i.e., two vectors of one element each).

LEMMA 5.1. *Let $\mathbf{P}$ be an NCD Markov chain with N blocks that is not exactly lumpable. Let $\mathbf{C}^s$ be the the $N \times N$ principal submatrix of the quasi-lumped chain $\mathbf{K}^s$ corresponding to $\mathbf{P}$ in* (3.6). *Then $\mathbf{C}^s$ has entries that satisfy*

$$0 \leq c^s_{i,j} \leq \min\left(\mathbf{P}_{i,j}\boldsymbol{e}\right) \ll 1 \qquad \forall i \neq j, \tag{5.1}$$

$$0 \ll \min(\mathbf{P}_{i,i}\boldsymbol{e}) \leq c^s_{i,i} < 1 \qquad \forall i. \tag{5.2}$$

Proof. Once again introduce $k_{i,j} = \min(\mathbf{P}_{i,j}\boldsymbol{e})$. Now observe that

$$0 \leq k_{i,j} \ll 1 \qquad \forall i \neq j,$$

$$0 \ll k_{i,i} < 1 \qquad \forall i$$

are direct consequences of the following properties of NCD Markov chains [10].

- For off-diagonal blocks,

$$\mathbf{o} \leq \mathbf{P}_{i,j}\boldsymbol{e} \ll \boldsymbol{e} \qquad \forall i \neq j.$$

- For diagonal blocks,

$$\mathbf{o} \ll \mathbf{P}_{i,i}\boldsymbol{e} \leq \boldsymbol{e} \qquad \forall i$$

with the condition that $\mathbf{P}_{i,i}\boldsymbol{e} \neq \boldsymbol{e}$ (since $\mathbf{P}$ is irreducible by definition).

Now inspect the off-diagonal blocks in $\mathbf{P}^-$ (see (3.1)) given by (3.2). If $\mathbf{P}_{i,j}$ has equal row sums (i.e., $\mathbf{P}_{i,j}\boldsymbol{e} = k_{i,j}\boldsymbol{e}$), then $\mathbf{P}^-_{i,j} = \mathbf{P}_{i,j}$. Otherwise, $\mathbf{P}^-_{i,j} = \mathbf{P}_{i,j} - \mathbf{P}^\epsilon_{i,j}$, where $(\mathbf{P}_{i,j} - \mathbf{P}^\epsilon_{i,j})\boldsymbol{e} = k_{i,j}\boldsymbol{e}$. In all cases, $\mathbf{P}^-_{i,j}\boldsymbol{e} = k_{i,j}\boldsymbol{e}$. As for the diagonal blocks in $\mathbf{P}^-$, each diagonal block is equal to its counterpart in $\mathbf{P}$. Using (3.3), a new matrix $\mathbf{P}^s$ is formed. The only blocks (possibly) prohibiting lumpability in $\mathbf{P}^s$ are those diagonal blocks with unequal row sums. In other words, for $\mathbf{P}^s$ to be lumpable, each diagonal block i for which $\min(\mathbf{P}_{i,i}\boldsymbol{e}) \neq \max(\mathbf{P}_{i,i}\boldsymbol{e})$ (i.e., $\max(\bar{\boldsymbol{y}}_i) \neq \min(\bar{\boldsymbol{y}}_i)$) needs to be adjusted. The adjustment in $\mathbf{P}^s_{i,i}$ may be performed by increasing some elements in $\bar{\boldsymbol{y}}_i$ so as to make each element in $\bar{\boldsymbol{y}}_i$ equal to $\max(\bar{\boldsymbol{y}}_i)$ and decreasing the corresponding diagonal element in $\mathbf{P}^s_{i,i}$. The intended effect is to have $\mathbf{P}^s_{i,i}\boldsymbol{e} = k_{i,i}\boldsymbol{e}$. As a result of this diagonal adjustment, one obtains a new $\mathbf{P}^s$ which may or may not have negative elements along the diagonal. These two cases should be analyzed in turn.

(i) There are no negative elements along the diagonal of $\mathbf{P}^s$. Hence, the scaling in (3.5) need not be performed. In this case, $\tilde{\mathbf{P}}^s = \mathbf{P}^s$ (i.e., $\alpha = 1$ in (3.5)) and $\tilde{\mathbf{P}}^s$ may be quasi-lumped to form $\mathbf{K}^s$. The effect of quasi-lumping $\tilde{\mathbf{P}}^s$ is to have

$$k^s_{i,j} = k_{i,j} \qquad \forall i, j \in \{1, 2, \ldots, N\}. \tag{5.3}$$

(ii) There are one or more negative elements along the diagonal of $\mathbf{P}^s$. The scaling in (3.5) is performed. In this case, $\tilde{\mathbf{P}}^s = \alpha\mathbf{P}^s + (1-\alpha)\mathbf{I}$, where $0 < \alpha < 1$. The scalar α may be chosen so that the largest negative element in magnitude along the diagonal of $\mathbf{P}^s$ becomes zero after the scaling operation and $\tilde{\mathbf{P}}^s_{i,i} \geq \mathbf{0}\ \forall i$.

$$\tilde{\mathbf{P}}^s_{i,j}\boldsymbol{e} = \alpha\mathbf{P}^s_{i,j}\boldsymbol{e} \ \Rightarrow\ k^s_{i,j} = \alpha k_{i,j} \ \Rightarrow\ 0 \leq k^s_{i,j} \leq k_{i,j} \qquad \forall i \neq j, \tag{5.4}$$

(5.5)

$$\tilde{\mathbf{P}}^s_{i,i}\boldsymbol{e} = \alpha\mathbf{P}^s_{i,i}\boldsymbol{e} + (1-\alpha)\boldsymbol{e} \ \Rightarrow\ k^s_{i,i} = \alpha k_{i,i} + (1-\alpha) = \begin{cases} k_{i,i} + (1-\alpha)(1-k_{i,i}) \\ 1 - \alpha(1-k_{i,i}) \end{cases}$$

$$\Rightarrow\ k_{i,i} < k^s_{i,i} < 1 \qquad \forall i.$$

Combining the above two cases with the properties of NCD chains and noticing that $\mathbf{C}^s$ is the $N \times N$ principal submatrix of $\mathbf{K}^s$, one obtains the statement in the lemma. Once again it must be remarked that if (3.5) is not performed, then case (i) applies and $c^s_{i,j} = \min(\mathbf{P}_{i,j}\boldsymbol{e})\ \forall i,j$. ∎

LEMMA 5.2. *Let* $\mathbf{P}$ *be an NCD Markov chain with* N *blocks that is not exactly lumpable. Let* $\mathbf{C}^s$ *be the* $N \times N$ *principal submatrix of the quasi-lumped chain* $\mathbf{K}^s$ *corresponding to* $\mathbf{P}$ *in* (3.6). *Then* $\mathbf{C}^s$ *has entries that satisfy*

$$\sum_j c^s_{i,j} \leq 1 \qquad \forall i \tag{5.6}$$

with strict inequality for at least one i.

Proof. For the case in which scaling is not performed, the proof is straightforward and follows from (5.3):

$$\sum_j c^s_{i,j} = \sum_j k^s_{i,j} = \sum_j k_{i,j} = \sum_j \min(\mathbf{P}_{i,j}\boldsymbol{e}) \leq 1 \qquad \forall i.$$

The fact that there is strict inequality for at least one row of blocks is a consequence of $\mathbf{P}$ not being exactly lumpable. That is, there is at least one row of blocks in $\mathbf{P}$ in which one of the blocks has unequal row sums; otherwise, $\mathbf{P}$ would be exactly lumpable. When scaling is performed, one obtains

$$\begin{aligned}\sum_j c^s_{i,j} &= k^s_{i,i} + \sum_{j\neq i} k^s_{i,j} = 1 - \alpha(1 - k_{i,i}) + \alpha \sum_{j\neq i} k_{i,j} = 1 - \alpha + \alpha \sum_j k_{i,j} \\ &= 1 - \alpha + \alpha \sum_j \min(\mathbf{P}_{i,j}\boldsymbol{e}) \leq 1 \qquad \forall i\end{aligned}$$

from (5.4) and (5.5). The strict inequality for at least one i stems from the same reason. ∎

The following theorem summarizes the properties of $\mathbf{C}^s$.

THEOREM 5.3. *Let* $\mathbf{P}$ *be an NCD Markov chain with* N *blocks and coupling matrix* $\mathbf{C}$. *Assume that* $\mathbf{P}$ *is not exactly lumpable. Let* $\mathbf{C}^s$ *be the* $N \times N$ *principal submatrix of the quasi-lumped chain* $\mathbf{K}^s$ *corresponding to* $\mathbf{P}$ *in* (3.6). *Then*

(i) $\mathbf{C}^s$ *is nonnegative*;

(ii) $\mathbf{C}^s$ *is row diagonally dominant*;

(iii) $\mathbf{C}^s$ *may be reducible (although* $\mathbf{C}$ *is irreducible)*;

(iv) *if* $\mathbf{C}^s$ *is irreducible or each row of blocks in* $\mathbf{P}$ *is not exactly lumpable, then* $\mathbf{I} - \mathbf{C}^s$ *is a nonsingular M-matrix*;

(v) *if the scaling in* (3.5) *is not performed,* $\mathbf{C}^s \leq \mathbf{C}$;

(vi) *if the scaling in* (3.5) *is not performed and for some* i,j $\mathbf{P}_{i,j}$ *has equal row sums, then* $c_{i,j} = c^s_{i,j}$.

Proof. Parts (i) and (ii) follow directly from Lemma 5.1. Although the coupling matrix of an NCD Markov chain is irreducible, $\mathbf{C}^s$ may very well be a reducible matrix. The reason for this is implicit in equation (5.1). For a given diagonal element of $\mathbf{C}^s$, all off-diagonal elements in the same row may be zero. This is a sufficient condition and happens, for instance, if $\min(\mathbf{P}_{i,j}\boldsymbol{e}) = 0\ \forall i \neq j$ for a given i, and part (iii) follows. Note that it is also possible for $\mathbf{C}^s$ to be an irreducible matrix. For part (iv), let $\mathbf{A} = \mathbf{I} - \mathbf{C}^s$. To prove that $\mathbf{A}$ is a nonsingular M-matrix [1], the following properties need to be shown (see [9, pp. 531–532]):

1. $a_{i,i} > 0\ \forall i$ and $a_{i,j} \leq 0\ \forall i \neq j$.

2. A is irreducible and $a_{i,i} \geq \sum_{j \neq i} |a_{i,j}| \ \forall i$ with strict inequality for at least one i, or $a_{i,i} > \sum_{j \neq i} |a_{i,j}| \ \forall i$.

Now,

$$0 < a_{i,i} \ll 1 \ \forall i \quad \text{and} \quad -1 \ll a_{i,j} \leq 0 \ \forall i \neq j$$

follow directly from Lemma 5.1; hence, the first property is verified. The second property amounts to showing that $\sum_j c^s_{i,j} < 1 \, \forall i$. As indicated in Lemma 5.2, this is not true in general. However, if $\mathbf{C}^s$ is irreducible, then so is $\mathbf{A}$, and the second property is also satisfied due to Lemma 5.2. On the other hand, if strict inequality holds for each row of $\mathbf{C}^s$ in Lemma 5.2 (i.e., $\mathbf{A}$ is strictly row diagonally dominant), the irreducibility assumption for $\mathbf{C}^s$ may be relaxed and the second property is once again satisfied. Note that this is the case if each row of blocks in $\mathbf{P}$ possesses at least one block with unequal row sums, and therefore it is quite likely to happen. Finally, the nonsingularity is a direct consequence of condition (I_{29}) on p. 136 of [1]. Part (v) follows from Lemma 5.1. A sufficient condition for $\mathbf{C}^s \leq \mathbf{C}$ to be true is for $\mathbf{P}$ to be diagonally dominant or for $\mathbf{P}$ to have diagonal elements larger than the degree of coupling. Part (vi) may be shown by noticing that $c^s_{i,j} = \min(\mathbf{P}_{i,j}\boldsymbol{e})$ if $\mathbf{P}_{i,j}$ has equal row sums and scaling is not performed. Hence,

$$c_{i,j} = \frac{\boldsymbol{\pi}_i}{\|\boldsymbol{\pi}_i\|_1}\mathbf{P}_{i,j}\boldsymbol{e} = \frac{\boldsymbol{\pi}_i}{\|\boldsymbol{\pi}_i\|_1}c^s_{i,j}\boldsymbol{e} = c^s_{i,j}\frac{\boldsymbol{\pi}_i\boldsymbol{e}}{\|\boldsymbol{\pi}_i\|_1} = c^s_{i,j} \qquad \forall i,j. \qquad \square$$

COROLLARY 5.4. *Let $\mathbf{P}$ be an NCD Markov chain with N blocks and coupling matrix $\mathbf{C}$. Then $\mathbf{C}^l$ with entries*

$$c^l_{i,j} = \min(\mathbf{P}_{i,j}\boldsymbol{e}) \qquad \forall i,j \tag{5.7}$$

is a nonnegative, lower-bounding matrix for $\mathbf{C}$ and $\mathbf{C}^u$ with entries

$$c^u_{i,j} = \max(\mathbf{P}_{i,j}\boldsymbol{e}) = \|\mathbf{P}_{i,j}\|_\infty \qquad \forall i,j \tag{5.8}$$

is a nonnegative, upper-bounding matrix for $\mathbf{C}$.

That $\mathbf{C} \leq \mathbf{C}^u$ follows from

$$c_{i,j} = \frac{\boldsymbol{\pi}_i}{\|\boldsymbol{\pi}_i\|_1}\mathbf{P}_{i,j}\boldsymbol{e} \leq \max(\mathbf{P}_{i,j}\boldsymbol{e}) \qquad \forall i,j,$$

where $\boldsymbol{\pi}_i/\|\boldsymbol{\pi}_i\|_1$ is a probability vector. Also note that $\mathbf{C}^u$ is irreducible because $\mathbf{P}$ is irreducible, whereas an analogous statement is not valid for $\mathbf{C}^l$.

Returning to the question posed at the beginning of this section, the answer is no, $\mathbf{C}^s$ is not necessarily a lower-bounding matrix for $\mathbf{C}$, but $\mathbf{C}^l$ is. Nevertheless, for the Courtois example $\mathbf{C}^s = \mathbf{C}^l$, and $\mathbf{C}^s$ turns out to be a lower-bounding matrix for $\mathbf{C}$. Note that it is possible to subtract a slack probability mass from some other element (rather than the diagonal element) in the diagonal block and avoid the scaling in equation (3.5) (see the third step of construction in section 3) to have $\mathbf{C}^s = \mathbf{C}^l$. We use the definition of quasi lumpability in [5] to be consistent in terminology. The next section investigates the relation between the nonzero structure of a substochastic lower-bounding matrix for a given Markov chain and the nature of lower and upper bounds obtained on the chain's stationary probabilities.

6. Significance of the structure of lower-bounding matrices. Given an irreducible Markov chain $\mathbf{P}$ and a substochastic lower-bounding matrix $\mathbf{P}^*$ (i.e., $\mathbf{0} \leq \mathbf{P}^* \leq \mathbf{P},\ \mathbf{P}^* \neq \mathbf{0}$), one can use Courtois and Semal's technique and compute lower and upper bounds on the stationary probabilities of $\mathbf{P}$. The question of interest is the following. What, if any, is the relation between the nonzero structure of $\mathbf{P}^*$ and the bounds obtained? Analogously, the same question may be posed for the coupling matrix of an NCD Markov chain that is not exactly lumpable and a nonnegative lower-bounding coupling matrix $\mathbf{C}^*$ (such as $\mathbf{C}^l$ of (5.7)) (i.e., $\mathbf{0} \leq \mathbf{C}^* \leq \mathbf{C},\ \mathbf{C}^* \neq \mathbf{0}$). In order to avoid introducing new symbols and complicating the terminology further, the equivalent second question is considered. That $\mathbf{C}^l$ and the like have weighty diagonals is immaterial in the theory developed.

Observe that $\mathbf{C}^* \geq 0$, $c^*_{i,i} \neq 0\ \forall i$, and $\mathbf{C}^*\boldsymbol{e} \neq \boldsymbol{e}$ for the matrices of interest by definition. The principles that govern the solution of the systems

$$\boldsymbol{z}_i\mathbf{K}^*_i = \boldsymbol{z}_i \qquad \sum_{j=1}^{N} z_{i,j} = 1 \qquad \forall s_i \in \mathcal{S}^* = \{s_1, s_2, \ldots, s_N\}, \tag{6.1}$$

where

$$\mathbf{K}^*_i = \begin{pmatrix} \mathbf{C}^* & \boldsymbol{e} - \mathbf{C}^*\boldsymbol{e} \\ \boldsymbol{e}_i^T & 0 \end{pmatrix}, \tag{6.2}$$

are established next. Here $\mathbf{K}^*_i$ is a stochastic matrix (i.e., $\mathbf{K}^*_i\boldsymbol{e} = \boldsymbol{e}$), $\boldsymbol{z}_i$ is a probability vector (i.e., the ith row of the stochastic matrix $\mathbf{Z}$), $\mathcal{S}^*$ represents the states of the lower-bounding nonnegative (coupling) matrix, and $\boldsymbol{e}_i$ denotes the ith column of the identity matrix.

The discussion that follows refers to essential and nonessential (i.e., transient) states and to the concept of reducibility in nonnegative square matrices, as presented in pages 25–26 of [16]. Furthermore, for simplicity it is assumed that $\mathbf{C}^*$ is already in the normal form of a reducible (i.e., decomposable) nonnegative matrix. However, that $\mathbf{C}^*$ is in reducible normal form should not be understood to mean $\mathbf{C}^*$ is reducible.

Following the terminology in [16], let K denote the number of mutually disjoint irreducible subsets of states in $\mathbf{C}^*$. Let these subsets be represented by $\mathcal{S}^{ir}_1, \mathcal{S}^{ir}_2, \ldots, \mathcal{S}^{ir}_K$. Note that $\mathcal{S}^{ir}_i \bigcap \mathcal{S}^{ir}_j = \emptyset\ \forall i \neq j$. In any case, the states in $\mathcal{S}^{ir}$ $(= \bigcup_i \mathcal{S}^{ir}_i)$ are referred to as essential states. If $\mathcal{S}^{ir} = \mathcal{S}^*$, there would be no transient states in $\mathbf{C}^*$. Moreover, if $K = 1$, $\mathbf{C}^*$ would be irreducible; else it could be decomposed into K mutually disjoint irreducible subsets of states. Hereafter, the possibility of having a stochastic transition probability submatrix (as part of $\mathbf{C}^*$) corresponding to any irreducible subset of states is overruled. That is, for each irreducible subset of states, the extra column in $\mathbf{K}^*_i$ has at least one nonzero element. If not, the irreducible subset of states for which this property does not hold may be extracted from $\mathbf{C}^*$ and analyzed separately. On the other hand, if $\mathcal{S}^{ir} \neq \mathcal{S}^*$, there would be transient states in $\mathbf{C}^*$. Similarly, let $\mathcal{S}^{tr}_1, \mathcal{S}^{tr}_2, \ldots, \mathcal{S}^{tr}_M$ represent the transient subsets of states, where M is the number of transient subsets of states in $\mathbf{C}^*$ subject to the constraints $\mathcal{S}^{tr}_i \bigcap \mathcal{S}^{tr}_j = \emptyset$ $\forall i \neq j$. Moreover, the mutually disjoint transient subsets of states should be ordered so that there are no transitions from $\mathcal{S}^{tr}_i$ to $\mathcal{S}^{tr}_j$ in $\mathcal{S}^{tr}(= \bigcup_i \mathcal{S}^{tr}_i)\ \forall i < j$. However, there must be a transition from a given $\mathcal{S}^{tr}_i$ to at least one $\mathcal{S}^{tr}_k$ for $1 \leq k < i \leq M$ or to at least one $\mathcal{S}^{ir}_l$ for $1 \leq l \leq K$.

The following 9×9 lower-bounding nonnegative (coupling) matrix for an (NCD) Markov chain demonstrates the concepts introduced in this section.

Example 6.1. Let

$$\mathbf{C}^* = \left(\begin{array}{c|cc|ccc|cc|c} 0.999 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0.995 & 0.005 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.002 & 0.997 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0.998 & 0.001 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.997 & 0.003 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.002 & 0 & 0.998 & 0 & 0 & 0 \\ \hline 0 & 0.001 & 0 & 0 & 0 & 0 & 0.997 & 0.002 & 0 \\ 0 & 0.002 & 0.002 & 0 & 0 & 0 & 0.001 & 0.995 & 0 \\ \hline 0.001 & 0 & 0 & 0 & 0 & 0.001 & 0 & 0.001 & 0.996 \end{array}\right).$$

For this matrix, $\mathcal{S}^{ir} = \{s_1, s_2, \ldots, s_6\}$ and $\mathcal{S}^{tr} = \{s_7, s_8, s_9\}$ with $K = 3$, $M = 2$, $\mathcal{S}_1^{ir} = \{s_1\}$, $\mathcal{S}_2^{ir} = \{s_2, s_3\}$, $\mathcal{S}_3^{ir} = \{s_4, s_5, s_6\}$, $\mathcal{S}_1^{tr} = \{s_7, s_8\}$, $\mathcal{S}_2^{tr} = \{s_9\}$. Since $\mathbf{C}^*$ is in reducible normal form, each diagonal block in $\mathbf{C}^*$ is (and should be) irreducible. By the same token, the first transient subset of states, $\mathcal{S}_1^{tr}$, always has a transition to an irreducible subset of states from which the extra state is accessible. Therefore, by induction all transient subsets of states can access the extra state. For this example, the nonzero structure of $\mathbf{Z}$ in (6.1), (6.2) is given by the following matrix in which an X represents a nonzero entry:

$$\left(\begin{array}{c|cc|ccc|cc|c} \mathrm{X} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & \mathrm{X} & \mathrm{X} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathrm{X} & \mathrm{X} & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & \mathrm{X} & \mathrm{X} & \mathrm{X} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{X} & \mathrm{X} & \mathrm{X} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{X} & \mathrm{X} & \mathrm{X} & 0 & 0 & 0 \\ \hline 0 & \mathrm{X} & \mathrm{X} & 0 & 0 & 0 & \mathrm{X} & \mathrm{X} & 0 \\ 0 & \mathrm{X} & \mathrm{X} & 0 & 0 & 0 & \mathrm{X} & \mathrm{X} & 0 \\ \hline \mathrm{X} & \mathrm{X} & \mathrm{X} & \mathrm{X} & \mathrm{X} & \mathrm{X} & \mathrm{X} & \mathrm{X} & \mathrm{X} \end{array}\right).$$

The following theorems summarize these observations, enabling one to forecast the nonzero structure of $\mathbf{Z}$ for a given $\mathbf{C}^*$. It should be emphasized once more that each irreducible subset of states in the lower-bounding nonnegative matrices of interest should have a transition to the extra state and that the original Markov chain should not be exactly lumpable. Under these conditions, one may state the following theorems, which are valid a fortiori for an NCD Markov chain with coupling matrix $\mathbf{C}$ such that $\mathbf{C}^* \leq \mathbf{C}$ and $\mathbf{C}^*$ is substochastic.

THEOREM 6.2. *Let $\mathbf{C}^*$ be a substochastic matrix. If $\mathbf{C}^*$ is irreducible, then $\mathbf{Z}$ given by* (6.1), (6.2) *is positive.*

Proof. Since $\mathbf{C}^*\boldsymbol{e} \neq \boldsymbol{e}$, there is at least one row in $\mathbf{C}^*$, say k, for which $(\boldsymbol{e} - \mathbf{C}^*\boldsymbol{e})_k > 0$. All states in $\mathbf{C}^*$ form a single communicating class and the extra state in $\mathbf{K}_i^*$ (see (6.2)) is accessible from at least one of the states in $\mathbf{C}^*$. Hence, $\mathbf{K}_i^*$ is irreducible for each i, and the theorem follows. □

Note that when $\mathbf{C}^*$ is irreducible, $\mathcal{S}^{ir} = \mathcal{S}^*$, $K = 1$, and there are no transient states in $\mathcal{S}^*$. Furthermore, under the stated conditions $\mathbf{I} - \mathbf{C}^*$ is a nonsingular M-matrix.

In the statement of the following theorem, a substochastic state means a state for which the corresponding row sum is less than one.

THEOREM 6.3. *Let $\mathbf{C}^*$ be a substochastic matrix, and let $\mathcal{S}^* = \mathcal{S}^{ir} \bigcup \mathcal{S}^{tr}$, $\mathcal{S}^{ir} = \bigcup_{i=1}^{K} \mathcal{S}_i^{ir}$, $\mathcal{S}^{tr} = \bigcup_{i=1}^{M} \mathcal{S}_i^{tr}$ be the state space partition of $\mathbf{C}^*$, where K is the number of*

disjoint irreducible subsets of states and M is the number of disjoint transient subsets of states. If $\mathbf{C}^*$ is reducible and each irreducible subset of states in $\mathbf{C}^*$ has at least one substochastic state, then

(i) *if s_i is an essential state and $s_i \in \mathcal{S}_k^{ir}$ for some k, then $z_{i,j} > 0$ for all $s_j \in \mathcal{S}_k^{ir}$ and $z_{i,j} = 0$ for all $s_j \notin \mathcal{S}_k^{ir}$;*

(ii) *if s_i is a transient state and $s_i \in \mathcal{S}_k^{tr}$ for some k, then $z_{i,j} > 0$ for all $s_j \in (\mathcal{S}_k^{tr}$ and states accessible from $\mathcal{S}_k^{tr})$; otherwise $z_{i,j} = 0$.*

Proof. Part (i) follows from the fact that the extra state is accessible from $\mathcal{S}_k^{ir}$, of which s_i is a member, and the last row of $\mathbf{K}_i^*$ has a one at the ith column position in (6.2), thereby making $\mathcal{S}_k^{ir}$ with the extra state an irreducible stochastic submatrix in $\mathbf{K}_i^*$. Hence, $\boldsymbol{z}_i$ in (6.1) has nonzero entries only in locations corresponding to the members of $\mathcal{S}_k^{ir}$. Part (ii) follows from the fact that the extra state is accessible from $\mathcal{S}_k^{tr}$ (of which s_i is a member) and all other subsets of states accessible from $\mathcal{S}_k^{tr}$. Hence, states in $\mathcal{S}_k^{tr}$ and states accessible from $\mathcal{S}_k^{tr}$ together with the extra state form an irreducible stochastic submatrix in $\mathbf{K}_i^*$. Again, $\boldsymbol{z}_i$ has nonzero entries only in locations corresponding to the members of $\mathcal{S}_k^{tr}$ and other states they access. ∎

COROLLARY 6.4. *If the substochastic matrix $\mathbf{C}^*$ is reducible and the kth irreducible subset of states $\mathcal{S}_k^{ir}$ is a singleton with a substochastic state (i.e., $\mathcal{S}_k^{ir} = \{s_i\},\ s_i \in \mathcal{S}^{ir}$), then $z_{i,j} = \delta_{i,j}$.*

Corollary 6.4 helps to identify those states for which the lower and upper bounds obtained by Courtois and Semal's technique will be 0 and 1, respectively. Such states do not contribute to the tightening of the bounds of other states. Hence, if these states are identified in advance, they may be extracted from the lower-bounding matrix, thereby reducing the size of the systems to be solved in (6.1) and (6.2).

Before stating the next corollary, we recall the definition of a reachability (or accessibility) matrix. The reachability matrix of a square matrix is constructed as follows. First, the given square matrix is represented as a directed graph. The graph must have a directed arc for each nonzero entry in the original matrix. Then a new matrix is formed whose i,jth entry is a one (zero) if and only if state j is accessible (inaccessible) from state i on the directed graph. The newly formed matrix is the reachability matrix corresponding to the original square matrix.

COROLLARY 6.5. *If each irreducible subset of states in the substochastic matrix $\mathbf{C}^*$ has at least one substochastic state, then the nonzero structure of $\mathbf{Z}$ in* (6.1), (6.2) *is identical to the nonzero structure of the reachability matrix of $\mathbf{C}^*$.*

Corollary 6.5 helps one to forecast the nonzero structure of $\mathbf{Z}$ by inspecting the nonzero structure of the lower-bounding matrix; that is, one does not need to solve N systems to find out what the nonzero structure of $\mathbf{Z}$ looks like.

A result of Theorem 6.3 and Corollaries 6.4 and 6.5 (with (4.2) and (4.3)) is that a reducible lower-bounding nonnegative matrix gives lower (upper) bounds of zero (one) for various stationary probabilities of the coupling matrix and therefore indirectly causes other stationary probabilities to be loosely bounded. In conclusion, reducible lower-bounding nonnegative matrices should be avoided whenever possible.

7. Conclusion. This paper shows that NCD Markov chains are quasi-lumpable (if not lumpable). In most cases, $\mathbf{C}^s$, the $N \times N$ principal submatrix of the quasi-lumped chain turns out to be a lower-bounding coupling matrix for an NCD chain with N NCD blocks. When $\mathbf{C}^s$ is a lower-bounding coupling matrix, it may used to compute lower and upper bounds for the stationary probabilities of the NCD blocks. If $\mathbf{C}^s$ is not a lower-bounding coupling matrix, $\mathbf{C}^l$, which is guaranteed to be a lower-bounding coupling matrix, may be used instead. Bounding the station-

ary probabilities of NCD blocks from below and from above amounts to solving at most N, $(N+1) \times (N+1)$, systems. These linear systems differ only in the last row. Therefore, only one **LU** decomposition needs to be performed. Assuming that the transposed systems of equations are solved, the upper-triangular matrices will be different in the last columns only. Hence, the last column in each of these systems needs to be treated separately during the triangularization phase. Thereafter, all back substitutions may be performed in parallel. Consequently, a solution method such as Gaussian elimination has a time complexity of $O(N^3)$ in the computation of the bounds.

If the NCD Markov chain is sparse with symmetries in its nonzero structure, it is quite likely that some elements of the unknown vector $\boldsymbol{x}$ in the quasi-lumped chain will turn out to be zero, thus tightening the bounds further as in the Courtois matrix. The more information one has regarding the distribution of the probability mass in $\boldsymbol{x}^T$, the tighter the lower and upper bounds become. In fact, there is a distribution $\boldsymbol{x}^T$ which gives the stationary probability of being in each NCD block exactly to working precision. However, although ϵ is always less than or equal to the degree of coupling of the NCD Markov chain, the lower-bounding nonnegative coupling matrix will have diagonal elements close to one, and it seems that the bounds obtained by the procedure generally will not be tight. The ill-conditioned nature of NCD Markov chains is once again noticed, but this time from a different perspective.

Furthermore, when choosing lower-bounding nonnegative matrices for Markov chains, one should be on the lookout for irreducible matrices. Reducible matrices should be avoided whenever possible because they provide lower (upper) bounds of zero (one) for various stationary probabilities, thereby indirectly causing other stationary probabilities to be loosely bounded.

Acknowledgments. The authors wish to thank the referees for their remarks which led to improvements in the manuscript.

REFERENCES

[1] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, PA, 1994.

[2] W. L. Cao and W. J. Stewart, *Iterative aggregation/disaggregation techniques for nearly uncoupled Markov chains*, J. Assoc. Comput. Mach., 32 (1985), pp. 702–719.

[3] P.-J. Courtois, *Decomposability: Queueing and Computer System Applications*, Academic Press, New York, 1977.

[4] P.-J. Courtois and P. Semal, *Bounds for the positive eigenvectors of nonnegative matrices and for their approximations by decomposition*, J. Assoc. Comput. Mach., 31 (1984), pp. 804–825.

[5] G. Franceschinis and R. R. Muntz, *Bounds for quasi-lumpable Markov chains*, Performance Evaluation, 20 (1994), pp. 223–243.

[6] W. J. Harrod and R. J. Plemmons, *Comparison of some direct methods for computing the stationary distributions of Markov chains*, SIAM J. Sci. Comput., 5 (1984), pp. 453–469.

[7] J. R. Kemeny and J. L. Snell, *Finite Markov Chains*, Van Nostrand, New York, 1960.

[8] J. R. Koury, D. F. McAllister, and W. J. Stewart, *Iterative methods for computing stationary distributions of nearly completely decomposable Markov chains*, SIAM J. Alg. Disc. Meth., 5 (1984), pp. 164–186.

[9] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, Academic Press, New York, 1985.

[10] C. D. Meyer, *Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems*, SIAM Rev., 31 (1989), pp. 240–272.

[11] C. D. Meyer, *Sensitivity of the stationary distribution of a Markov chain*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 715–728.

[12] P. J. Schweitzer, *A survey of aggregation–disaggregation in large Markov chains*, in Numerical Solution of Markov Chains, W. J. Stewart, ed., Marcel Dekker, New York, 1991, pp. 63–88.

[13] P. Semal, *Analysis of Large Markov Models, Bounding Techniques and Applications*, Doctoral Thesis, Université Catholique de Louvain, Belgium, 1992.

[14] G. W. Stewart, W. J. Stewart and D. F. McAllister, *A two–stage iteration for solving nearly completely decomposable Markov chains*, in IMA Volumes in Mathematics and its Applications 60: Recent Advances in Iterative Methods, G. H. Golub, A. Greenbaum, and M. Luskin, eds. Springer–Verlag, New York, 1994, pp. 201–216.

[15] W. J. Stewart and W. Wu, *Numerical experiments with iteration and aggregation for Markov chains*, ORSA J. Comput., 4 (1992), pp. 336–350.

[16] W. J. Stewart, *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, Princeton, NJ, 1994.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, pp. 499–517, April 1997

017

PROBABILISTIC ANALYSIS OF GAUSSIAN ELIMINATION WITHOUT PIVOTING*

MAN-CHUNG YEUNG† AND TONY F. CHAN†

Abstract. The numerical instability of Gaussian elimination is proportional to the size of the L and U factors that it produces. The worst-case bounds are well known. For the case without pivoting, breakdowns can occur and it is not possible to provide a priori bounds for L and U. For the partial pivoting case, the worst-case bound is $O(2^m)$, where m is the size of the system. Yet these worst-case bounds are seldom achieved, and in particular Gaussian elimination with partial pivoting is extremely stable in practice. Surprisingly, there has been relatively little theoretical study of the "average" case behavior. The purpose of our paper is to provide a probabilistic analysis of the case without pivoting. The distribution we use for the entries of A is the normal distribution with mean 0 and unit variance. We first derive the distributions of the entries of L and U. Based on this, we prove that the probability of the occurrence of a pivot less than ϵ in magnitude is $O(\epsilon)$. We also prove that the probabilities $\text{Prob}(||U||_\infty/||A||_\infty > m^{2.5})$ and $\text{Prob}(||L||_\infty > m^3)$ decay algebraically to zero as m tends to infinity. Numerical experiments are presented to support the theoretical results.

Key words. Gaussian elimination, pivot, growth factor, density function

AMS subject classifications. 65F05, 65G05

PII. S0895479895291741

1. Introduction.

Gaussian elimination (GE) is the most common general method for solving an $m \times m$, square, dense, unstructured linear system $Ax = b$. Together with partial pivoting, the method is extremely stable in practice. However, this stability cannot be guaranteed. The worst-case examples are well known: without pivoting breakdowns can occur and even with partial pivoting the "growth factor" can be as large as $O(2^m)$ (and can occur in practical applications [6]). This has motivated the "average-case" analysis [11] of GE in order to explain its practical numerical stability. Surprisingly, there has been relatively few other studies on this topic in the literature. The purpose of our paper is to provide a rather complete analysis for the case without pivoting.

Theoretical studies about the numerical stability of GE have been made since the 1940s by a great number of authors, for example, Turing [13], von Neumann and Goldstine [14], [15], Wilkinson [16], [17], and so on. Recently, Trefethen and Schreiber [11] considered the average-case analysis. Among their many results, they observed that for many distributions of matrices the matrix elements after the first few steps of GE with (partial or complete) pivoting are approximately normally distributed. They also found that, for $m \leq 1024$, the average growth factor (normalized by the standard deviation of the initial matrix elements) is within a few percent of $m^{2/3}$ for the partial pivoting case and approximately $m^{1/2}$ for the complete pivoting case. After having performed more extensive experiments, Edelman and Mascarenhas [4] suggested that the growth factor in the partial pivoting case may grow more like $m^{1/2}$ than $m^{2/3}$.

Following Trefethen and Schreiber, we study the probability of small pivots and large growth factors in this paper. However, we will consider only the case without

* Received by the editors September 13, 1995; accepted for publication (in revised form) by G. Cybenko June 14, 1996. The authors were partially supported by NSF contract ASC-92-01266 and ONR contract ONR-N00014-92-J-1890.

http://www.siam.org/journals/simax/18-2/29174.html

† Department of Mathematics, University of California, Los Angeles, CA 90095-1555 (myeung@math.ucla.edu, chan@math.ucla.edu).

pivoting. We are doing so for three reasons. The first is quite obvious: the nonpivoting case is far easier to analyze than the pivoting case. In particular, we are able to derive in close form the density functions of the elements of the LU factors and probabilistic bounds for the occurrence of small pivots and the growth factors. The second reason is that with the advent of parallel computing there is more incentive to trade off the stability of partial pivoting for the higher performance of simpler but possibly less stable forms of GE, including no pivoting; see, for instance, [8, 9]. Finally, we are hoping that our results for GE without pivoting will be useful in the analysis of, as well as provide a basis of comparison for, the partial pivoting case.

Throughout the paper, we suppose $X \in R^{m\times m}$ is a random matrix with independent and identically distributed elements which are $N(0,1)$, the normal distribution with mean 0 and variance 1. This choice is motivated by the empirical results of Trefethen and Schreiber mentioned earlier. Matrices of this type have also been studied by Edelman [2], [3], who derived the expected singular values.

In sections 2 and 3, we derive the density functions of the entries of L and U, respectively, where $X = LU$, the LU factorization of X. In section 4, we prove that the probability of the occurrence of a pivot less than ϵ in magnitude is $O(\epsilon)$.[1] In section 5, we derive bounds on the probabilities of large growth factors. In particular, we prove that the probabilities $\text{Prob}(||U||_\infty/||A||_\infty > m^{2.5})$ and $\text{Prob}(||L||_\infty > m^3)$ decay algebraically to zero as m tends to infinity. Finally, we present experimental results in section 6. We observe that the probabilities $\text{Prob}(m \le ||L||_\infty < m^{1.5})$ and $\text{Prob}(m \le ||U||_\infty/||A||_\infty < m^{1.5})$ tend to one as m goes to infinity. This indicates that our theoretical bounds are not the tightest possible, but not too loose either.

2. Density function of u_{pq}. Let X be an $m \times m$ real matrix with independent and identically distributed elements from $N(0,1)$, to which we simply refer as "$X \sim \mathcal{N}_m(O, I)$." Let $X = LU$, where L is a unit lower triangular matrix and U is an upper triangular matrix, be the LU factorization of X.[2] The (p,q)th $(p \le q)$ entry u_{pq} of U and the entries of X have the following relation.

LEMMA 1. *Let $X = LU$ be the LU factorization of X. Then*

$$u_{pq} = x_{pq} - x_{p*}^T X_{p-1}^{-1} x_{*q},$$

where

$$x_{p*} = (x_{p1}, \ldots, x_{pp-1})^T,$$
$$x_{*q} = (x_{1q}, \ldots, x_{p-1q})^T,$$

and X_{p-1} is the $(p-1)\times(p-1)$ leading principal submatrix of X.

Proof. Permuting the pth and qth columns of X and U simultaneously on both sides of $X = LU$ and then comparing the corresponding blocks, we find that

$$\begin{bmatrix} X_{p-1} & x_{*q} \\ x_{p*}^T & x_{pq} \end{bmatrix} = \begin{bmatrix} L_{p-1} & 0 \\ l_{p*}^T & 1 \end{bmatrix} \begin{bmatrix} U_{p-1} & u_{*q} \\ 0 & u_{pq} \end{bmatrix},$$

where

$$l_{p*} = (l_{p1}, \ldots, l_{pp-1})^T,$$
$$u_{*q} = (u_{1q}, \ldots, u_{p-1q})^T,$$

[1] We note that Foster [5] has studied the probability of large diagonal elements in the QR factorization of a rectangular matrix A.

[2] Since they just form a set of measure 0, we ignore matrices for which GE fails.

and where L_{p-1} and U_{p-1} are the $(p-1)\times(p-1)$ leading principal submatrices of L and U, respectively. It follows that

$$X_{p-1} = L_{p-1}U_{p-1}, \qquad l_{p*}^T = x_{p*}^T U_{p-1}^{-1},$$
$$u_{*q} = L_{p-1}^{-1}x_{*q}, \qquad u_{pq} = x_{pq} - l_{p*}^T u_{*q},$$

and these imply the desired equation. □

Let H be a $(p-1)\times(p-1)$ orthogonal matrix, e.g., a Householder matrix, such that

$$x_{p*}^T H = (0,\ldots,0,s) \equiv \eta^T$$

with $s \geq 0$. Then

$$\begin{aligned} u_{pq} &= x_{pq} - \eta^T (X_{p-1}H)^{-1} x_{*q} \\ &\equiv x_{pq} - \eta^T Y^{-1} x_{*q}. \end{aligned}$$

It can be shown that the entries s, x_{pq}, x_{iq}, and $y_{ij},\ i,j=1,\ldots,p-1$, are mutually independent and all x_{pq}, x_{iq}, and $y_{ij},\ i,j=1,\ldots,p-1$, are $N(0,1)$ while s^2 is χ^2_{p-1}. The proof basically follows the approach in [10] and [12]. We now decompose Y as

$$Y = QR,$$

where Q is a $(p-1)\times(p-1)$ orthogonal matrix and R a $(p-1)\times(p-1)$ upper triangular matrix with positive diagonal elements. We then further have

$$\begin{aligned} u_{pq} &= x_{pq} - \eta^T R^{-1} Q^T x_{*q} \\ &\equiv x_{pq} - \eta^T R^{-1}\omega \\ &= x_{pq} - \frac{s w_{p-1}}{r_{p-1\,p-1}}. \end{aligned} \tag{1}$$

Again, the variables s, x_{pq}, w_i, and $r_{ij},\ i\leq j,\ i,j=1,\ldots,p-1$, are independent. s^2 is χ^2_{p-1} and r_{ii}^2 is $\chi^2_{p-i},\ i=1,\ldots,p-1$ and all others are $N(0,1)$. The proof basically follows the approach in [10] and [12].

Since the variables in the right-hand side of (1) are independent and their density functions are known, it is straightforward to determine the density function of u_{pq}.

THEOREM 1. *Suppose $X \sim \mathcal{N}_m(O,I)$ and let $X = LU$ be the LU factorization of X. Then the density function of the (p,q)th entry of U is*

$$f_{u_{pq}}(t) = \frac{\sqrt{2}}{\pi}\,\frac{\Gamma(p/2)}{\Gamma((p-1)/2)}\left(\sum_{i=0}^{\lfloor\frac{p-3}{2}\rfloor}\xi_{i,p}\,t^{-2i-2} + (-1)^{\lfloor\frac{p-1}{2}\rfloor}\,\zeta_p\,t^{-p+1}\exp\left(-\frac{1}{2}t^2\right)\,\phi_p(t)\right), \tag{2}$$

where

$$\xi_{i,p} = \begin{cases} (-1)^i \prod_{j=0}^{i-1} (p-2j-3), & i > 0, \\ 1, & i = 0, \end{cases}$$

$$\zeta_p = \begin{cases} (p-3)!!, & p > 3, \\ 1, & p = 2,\ 3, \end{cases}$$

$$\phi_p(t) = \left(\int_0^t \exp\left(\frac{1}{2}x^2\right) dx \right)^{p-1-2\lfloor (p-1)/2 \rfloor},$$

and where $-\infty < t < \infty$, $2 \le p \le q$.

Proof. Since the variables r^2_{p-1p-1} ($r_{p-1p-1} \ge 0$), s^2 ($s \ge 0$), w_{p-1}, and x_{pq} in (1) are χ^2_1, χ^2_{p-1}, $N(0,1)$, and $N(0,1)$, respectively, the density functions of r_{p-1p-1}, s, w_{p-1}, and x_{pq} are given as follows:

$$f_{r_{p-1p-1}}(t) = \begin{cases} \sqrt{\frac{2}{\pi}} \exp\left(-t^2/2\right), & t > 0, \\ 0, & t \le 0, \end{cases}$$

$$f_s(t) = \begin{cases} \frac{1}{2^{(p-3)/2}\Gamma((p-1)/2)} t^{p-2} \exp\left(-t^2/2\right), & t > 0, \\ 0, & t \le 0, \end{cases}$$

$$f_{w_{p-1}}(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-t^2/2\right),$$

and

$$f_{x_{pq}}(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-t^2/2\right).$$

Since r_{p-1p-1}, s, w_{p-1}, and x_{pq} are independent, their joint density function is given by

$$f(r,s,w,x) = f_{r_{p-1p-1}}(r)\, f_s(s)\, f_{w_{p-1}}(w)\, f_{x_{pq}}(x)$$

$$= \begin{cases} \tilde{c}\, s^{p-2} \exp\left(-\frac{1}{2}(s^2+r^2+w^2+x^2)\right), & r,\ s > 0, \\ 0 & \text{otherwise}, \end{cases}$$

where $\tilde{c} = \frac{1}{\pi^{3/2}2^{(p-2)/2}\Gamma((p-1)/2)}$. Thus, the distribution function $F_{u_{pq}}(\alpha)$ of u_{pq} is

$$\begin{aligned} F_{u_{pq}}(\alpha) &= \tilde{c}\iiiint_{u_{pq}\le\alpha} s^{p-2}\exp\left(-\frac{1}{2}(s^2 + r_{p-1p-1}^2 + w_{p-1}^2 + x_{pq}^2)\right) ds\, dr_{p-1p-1} dw_{p-1} dx_{pq} \\ &= \tilde{c}\iiiint_{x_{pq}-\frac{s\,w_{p-1}}{r_{p-1p-1}}\le\alpha} s^{p-2}\exp\left(-\frac{1}{2}(s^2 + r_{p-1p-1}^2 + w_{p-1}^2 + x_{pq}^2)\right) \\ &\quad \times ds\, dr_{p-1p-1} dw_{p-1} dx_{pq}\,. \end{aligned}$$

Using Lemma 3 in the Appendix, we can show that

$$f_{u_{pq}}(t) = \tilde{c}\int_0^\infty dx \int_{-\infty}^\infty \frac{x^{p-1}}{x^2+y^2}\exp\left(-\frac{1}{2}\left(x^2 + (y+t)^2\right)\right) dy, \tag{3}$$

which can be further reduced to (2) by Lemma 4 in the Appendix. □

3. Density function of l_{pq}. Similar to the derivation of the density function of u_{pq}, we first establish a relation between l_{pq} and the entries of X and then simplify it. Let $X = LU$ and $X^T = \tilde{L}\tilde{U}$ be the LU factorizations of X and X^T, respectively. Set $\tilde{D} = \mathrm{diag}(\tilde{u}_{11}, \ldots, \tilde{u}_{mm})$. Thus, $X^T = \tilde{L}\tilde{D}\tilde{D}^{-1}\tilde{U}$. So $X = (\tilde{D}^{-1}\tilde{U})^T(\tilde{L}\tilde{D})^T$. Note that $(\tilde{D}^{-1}\tilde{U})^T$ is unit lower triangular and $(\tilde{L}\tilde{D})^T$ upper triangular. By the uniqueness of the LU factorization of X, we have

$$L = \left(\tilde{D}^{-1}\tilde{U}\right)^T.$$

Hence

$$l_{pq} = \tilde{u}_{qp}/\tilde{u}_{qq}$$

for $1 \le q < p \le m$. By Lemma 1,

$$\tilde{u}_{qp} = x_{pq} - x_{*q}^T X_{q-1}^{-T} x_{p*}$$

and

$$\tilde{u}_{qq} = x_{qq} - x_{*q}^T X_{q-1}^{-T} x_{q*},$$

where

$$\begin{aligned} x_{p*} &= (x_{p1}, \ldots, x_{pq-1})^T, \\ x_{q*} &= (x_{q1}, \ldots, x_{qq-1})^T, \\ x_{*q} &= (x_{1q}, \ldots, x_{q-1q})^T, \end{aligned}$$

and X_{q-1} is the $(q-1)\times(q-1)$ leading principal submatrix of X. We now let H be a $(q-1)\times(q-1)$ orthogonal matrix such that

$$x_{*q}^T H = (0, \ldots, 0, s) \equiv \eta^T$$

with $s \ge 0$. Then

$$\begin{aligned} l_{pq} &= \frac{x_{pq} - \eta^T\left(X_{q-1}^T H\right)^{-1} x_{p*}}{x_{qq} - \eta^T\left(X_{q-1}^T H\right)^{-1} x_{q*}} \\ &\equiv \frac{x_{pq} - \eta^T Y^{-1} x_{p*}}{x_{qq} - \eta^T Y^{-1} x_{q*}}\,. \end{aligned}$$

As in the case of u_{pq} in section 2, all the entries in the above expression are mutually independent and s^2 is χ^2_{q-1} while others are $N(0,1)$. Let

$$Y = QR$$

be the QR factorization of Y where R has positive diagonal elements. Then the expression can be reduced to

$$\begin{aligned} l_{pq} &= \frac{x_{pq} - \eta^T R^{-1} Q^T x_{p*}}{x_{qq} - \eta^T R^{-1} Q^T x_{q*}} \\ &\equiv \frac{x_{pq} - \eta^T R^{-1} \omega}{x_{qq} - \eta^T R^{-1} \mu} \\ &= \frac{r_{q-1q-1} x_{pq} - s\omega_{q-1}}{r_{q-1q-1} x_{qq} - s\mu_{q-1}}. \end{aligned} \tag{4}$$

The entries x_{pq}, x_{qq}, ω_i, μ_i, and $r_{ij} (i < j)$ are $N(0,1)$ while s^2 is χ^2_{q-1} and r^2_{ii} is χ^2_{q-i}, where $i = 1, \ldots, q-1$, $j = 2, \ldots, q-1$. They are all independent.

THEOREM 2. *Suppose $X \sim \mathcal{N}_m(O, I)$ and let $X = LU$ be the LU factorization of X. Then the density function of the (p,q)th entry of L is*

$$f_{l_{pq}}(t) = \frac{1}{\pi} \frac{1}{1+t^2}, \tag{5}$$

where $-\infty < t < \infty$ and $1 \le q < p \le m$.

Proof. Suppose $q > 1$ and let $F_{l_{pq}}(\alpha)$ be the distribution function of l_{pq}. Since the joint density function of r_{q-1q-1}, x_{pq}, x_{qq}, ω_{q-1}, μ_{q-1}, and s is

$$f(r_{q-1q-1}, x_{pq}, x_{qq}, \omega_{q-1}, \mu_{q-1}, s)$$

$$= \begin{cases} \dfrac{1}{2^{q/2}\pi^{5/2}\Gamma((q-1)/2)} s^{q-2} \\ \times \exp\left(-\dfrac{1}{2}(r^2_{q-1q-1} + x^2_{pq} + x^2_{qq} + \omega^2_{q-1} + \mu^2_{q-1} + s^2)\right), & r_{q-1q-1},\ s > 0, \\ 0 & \text{otherwise}, \end{cases}$$

and since

$$F_{l_{pq}}(\alpha) = \int \cdots \int_{l_{pq} \le \alpha} f(r_{q-1q-1}, x_{pq}, x_{qq}, \omega_{q-1}, \mu_{q-1}, s) dr_{q-1q-1} dx_{pq} dx_{qq} d\omega_{q-1} d\mu_{q-1} ds,$$

(5) holds from (4) and Lemmas 5 and 6 in the Appendix. The case in which $q = 1$ is quite trivial if we notice that l_{p1} is the division of two $N(0,1)$ variables x_{p1} and x_{11}. □

REMARK. *In our private communications with him, Alan Edelman of MIT indicated to us that the proof of Theorem 2 can be greatly simplified by observing that every element of L is a ratio of two quantities x and y such that (x, y) is circularly symmetric and such a ratio has Cauchy distribution.*

4. Probability of small pivot. In practice, if one of the pivot elements u_{pp} is zero or smaller in magnitude than a preset tolerance ϵ, GE will fail. In this section, we describe the probability of the occurrence of such a situation. First, we give a bound on the density function $f_{u_{pq}}(t)$ of u_{pq}.

LEMMA 2.

$$\frac{1}{\pi\sqrt{2}}\,\frac{\Gamma(p/2)}{\Gamma((p+1)/2)}\,\exp\left(-\frac{t^2}{2}\right) \le f_{u_{pq}}(t) \le \frac{1}{\pi}\,\frac{\Gamma(p/2)}{\Gamma((p+1)/2)}\,\exp\left(\frac{t^2}{2}\right)$$

for $-\infty < t < \infty$ *and* $p \ge 2$.

Proof. From (3), we have

$$f_{u_{pq}}(t) = \tilde{c}\int_0^\infty dx \int_0^\infty \frac{x^{p-1}}{x^2+y^2}\left(\exp\left(-\frac{1}{2}\left(x^2+(y+t)^2\right)\right)\right.$$
$$\left.+\exp\left(-\frac{1}{2}\left(x^2+(y-t)^2\right)\right)\right) dy.$$

Letting $y = xz$, this can be written as

$$\begin{aligned} f_{u_{pq}}(t) &= \tilde{c}\int_0^\infty dx \int_0^\infty \frac{x^{p-2}}{1+z^2}\left(\exp\left(-\frac{1}{2}\left(x^2+(xz+t)^2\right)\right)\right. \\ &\qquad \left.+\exp\left(-\frac{1}{2}\left(x^2+(xz-t)^2\right)\right)\right) dz \\ &= \tilde{c}\int_0^\infty dx \int_0^\infty \frac{x^{p-2}}{1+z^2}\exp\left(-\frac{1}{2}\left((1+z^2)x^2+t^2\right)\right) \\ &\qquad \times\left(\exp\left(-xzt\right)+\exp\left(xzt\right)\right) dz. \end{aligned} \tag{6}$$

Since $\exp(\xi)+\exp(-\xi) \ge 2$, we have

$$\begin{aligned} f_{u_{pq}}(t) &\ge 2\tilde{c}\int_0^\infty dx \int_0^\infty \frac{x^{p-2}}{1+z^2}\exp\left(-\frac{1}{2}\left((1+z^2)x^2+t^2\right)\right) dz \\ &= 2\tilde{c}\,\exp\left(-\frac{1}{2}t^2\right)\int_0^\infty dz \int_0^\infty \frac{x^{p-2}}{1+z^2}\exp\left(-\frac{1}{2}(1+z^2)x^2\right) dx. \end{aligned}$$

Let $w = \frac{1}{2}(1+z^2)x^2$. Then

$$\begin{aligned} f_{u_{pq}}(t) &\ge \frac{\sqrt{2}}{\pi^{3/2}\Gamma((p-1)/2)} \\ &\qquad \times \exp\left(-\frac{1}{2}t^2\right)\int_0^\infty dz \int_0^\infty (1+z^2)^{-(p+1)/2} w^{(p-3)/2}\exp(-w)\, dw \\ &= \frac{1}{\pi\sqrt{2}}\,\frac{\Gamma(p/2)}{\Gamma((p+1)/2)}\,\exp\left(-\frac{1}{2}t^2\right). \end{aligned}$$

Moreover, from (6) we have

$$\begin{aligned} f_{u_{pq}}(t) &\le 2\tilde{c}\int_0^\infty dx \int_0^\infty \frac{x^{p-2}}{1+z^2}\exp\left(-\frac{1}{2}\left((1+z^2)x^2+t^2\right)\right)\exp\left(xz|t|\right)dz \\ &\le 2\tilde{c}\int_0^\infty dx \int_0^\infty \frac{x^{p-2}}{1+z^2}\exp\left(-\frac{1}{2}\left((1+z^2)x^2+t^2\right)\right)\exp\left(\frac{1}{2}\left(\frac{1}{2}(xz)^2+2t^2\right)\right)dz \\ &= 2\tilde{c}\exp\left(\frac{1}{2}t^2\right)\int_0^\infty dz\int_0^\infty \frac{x^{p-2}}{1+z^2}\exp\left(-\frac{1}{2}\left(1+\frac{1}{2}z^2\right)x^2\right)dx. \end{aligned}$$

Letting $u = \frac{1}{2}(1+\frac{1}{2}z^2)x^2$, we finally have

$$\begin{aligned} f_{u_{pq}}(t) &\le \frac{\sqrt{2}}{\pi^{3/2}\Gamma((p-1)/2)}\exp\left(\frac{1}{2}t^2\right)\int_0^\infty dz\int_0^\infty (1+z^2)^{-1}\left(1+\frac{1}{2}z^2\right)^{-(p-1)/2} \\ &\quad \times u^{(p-3)/2}\exp\left(-u\right)du \\ &= \frac{\sqrt{2}}{\pi^{3/2}}\exp\left(\frac{1}{2}t^2\right)\int_0^\infty (1+z^2)^{-1}\left(1+\frac{1}{2}z^2\right)^{-(p-1)/2}dz \\ &\le \frac{\sqrt{2}}{\pi^{3/2}}\exp\left(\frac{1}{2}t^2\right)\int_0^\infty \left(1+\frac{1}{2}z^2\right)^{-(p+1)/2}dz \\ &= \frac{2}{\pi^{3/2}}\exp\left(\frac{1}{2}t^2\right)\int_0^\infty (1+z^2)^{-(p+1)/2}dz \\ &= \frac{1}{\pi}\frac{\Gamma(p/2)}{\Gamma((p+1)/2)}\exp\left(\frac{1}{2}t^2\right). \qquad \square \end{aligned}$$

To make the statements below neatly, we use a shorthand notation. For given $\epsilon > 0$ and $1 \le p \le m$, we define

$$E_{p,\epsilon} = \{X \in R^{m\times m} |\, |u_{pp}| < \epsilon\}.$$

Then the event that at least one u_{pp} has $|u_{pp}| < \epsilon$ is naturally denoted by $\bigcup_{p=1}^m E_{p,\epsilon}$.

COROLLARY 1. *Suppose $X \sim \mathcal{N}_m(O, I)$ and let $X = LU$ be the LU factorization of X. Given $\epsilon > 0$ and $1 \le p \le m$,*

$$\text{Prob}(E_{p,\epsilon}) = \alpha_{p,\epsilon}\frac{\Gamma(p/2)}{\Gamma((p+1)/2)},$$

where $\frac{\sqrt{2}}{\pi}\int_0^\epsilon \exp(-\frac{1}{2}t^2)dt \le \alpha_{p,\epsilon} \le \frac{2}{\pi}\int_0^\epsilon \exp(\frac{1}{2}t^2)dt$.

Proof. For the case in which $p = 1$, it is sufficient to note that

$$\text{Prob}(E_{1,\epsilon}) = \text{Prob}(|x_{11}| < \epsilon) = \frac{1}{\sqrt{2\pi}}\int_{-\epsilon}^{\epsilon}\exp\left(-\frac{1}{2}t^2\right)dt.$$

Other cases are just the direct results of Lemma 2. $\square$

THEOREM 3. *Suppose $X \sim \mathcal{N}_m(O, I)$ and let $X = LU$ be the LU factorization of X. Then*

$$\text{Prob}\left(\bigcup_{p=1}^{m} E_{p,\epsilon}\right) \leq c(m)\,\epsilon\,\exp\left(\frac{1}{2}\epsilon^2\right), \tag{7}$$

where $c(m) = \frac{2}{\pi}\sum_{p=1}^{m} \frac{\Gamma(p/2)}{\Gamma((p+1)/2)}$.

Proof. Since $\text{Prob}(\bigcup_{p=1}^{m} E_{p,\epsilon}) \leq \sum_{p=1}^{m} \text{Prob}(E_{p,\epsilon})$, (7) follows by Corollary 1. ∎

The coefficient $c(m)$ of $\epsilon\,\exp\left(\frac{1}{2}\epsilon^2\right)$ is a rather slow-growing function of m. In fact, it is about 1800 even when $m = 10^6$. So, if ϵ is small enough, (7) will certainly give a satisfying bound for the desirable probability. Moreover, the right-hand side of (7) is approximately linear with ϵ for small ϵ.

5. Probability of large growth factor. When GE is performed on an $m \times m$ matrix A in floating point arithmetic, the computed LU factors $\hat{L}$ and $\hat{U}$ are produced. Then, by solving two corresponding triangular systems, we obtain the solution $\hat{x}$ to $Ax = b$. The computed solution $\hat{x}$ satisfies

$$(A + E)\,\hat{x} = b$$

with

$$|E| \leq m\boldsymbol{u}\left(3\,|A| + 5\,|\hat{L}|\,|\hat{U}|\right) + O(\boldsymbol{u}^2),$$

where $\boldsymbol{u}$ is the unit roundoff and where, for any matrix M, we use $|M|$ to denote the matrix obtained by taking the absolute value of the elements of M; see, for instance, [7, Theorem 3.3.2]. From this, it follows that

$$\|E\|_\infty \leq m\boldsymbol{u}\|A\|_\infty\left(3 + 5\|\hat{L}\|_\infty\frac{\|\hat{U}\|_\infty}{\|A\|_\infty}\right) + O(\boldsymbol{u}^2).$$

We define the growth factors ρ_L and ρ_U to be

$$\rho_L = \|L\|_\infty, \qquad \rho_U = \|U\|_\infty/\|A\|_\infty.$$

It is possible that ρ_L and ρ_U can be very large because small pivots can appear. The following theorem gives probabilistic bounds on the sizes of ρ_L and ρ_U.

THEOREM 4. *Suppose $X \sim \mathcal{N}_m(O, I)$ and let $X = LU$ be the LU factorization of X. Then there exist numbers $1 > b > 0$ and $c > 0$, independent of m, such that*

$$\text{Prob}(\rho_U > r) \leq \frac{c}{r}m^{5/2} + \min\left(\frac{c}{r}m^{7/2}, \frac{1}{m}\right) + b^m$$

and

$$\text{Prob}(\rho_L > r) \leq \frac{c}{r}m^3$$

for any $r \geq 1$.

Proof. We first claim that there exists a $c_1 > 0$, independent of m, such that

$$\text{Prob}(\|U\|_\infty > r) \leq \frac{c_1}{r}m^{7/2}. \tag{8}$$

In fact, by (3) we have

$$\begin{aligned}
f_{u_{pq}}(t) &= \tilde{c}\int_0^\infty dx \int_{|y+t|\geq|t|/2} \frac{x^{p-1}}{x^2+y^2}\exp\left(-\frac{1}{2}\left(x^2+(y+t)^2\right)\right)dy \\
&\quad +\tilde{c}\int_0^\infty dx \int_{|y+t|<|t|/2} \frac{x^{p-1}}{x^2+y^2}\exp\left(-\frac{1}{2}\left(x^2+(y+t)^2\right)\right)dy \\
&\leq \tilde{c}\int_0^\infty dx \int_{|y+t|\geq|t|/2} \frac{x^{p-1}}{x^2+y^2}\exp\left(-\frac{1}{2}\left(x^2+\frac{1}{4}t^2\right)\right)dy \\
&\quad +\tilde{c}\int_0^\infty dx \int_{|y+t|<|t|/2} \frac{x^{p-1}}{x^2+t^2/4}\exp\left(-\frac{1}{2}\left(x^2+(y+t)^2\right)\right)dy \\
&\leq \tilde{c}\exp\left(-\frac{1}{8}t^2\right)\int_0^\infty dx \int_{-\infty}^\infty \frac{x^{p-1}}{x^2+y^2}\exp\left(-\frac{1}{2}x^2\right)dy \\
&\quad +\frac{4\tilde{c}}{t^2}\int_0^\infty dx \int_{-\infty}^\infty x^{p-1}\exp\left(-\frac{1}{2}\left(x^2+(y+t)^2\right)\right)dy \\
&= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{8}t^2\right)+\frac{4\sqrt{2}}{\pi}\frac{\Gamma(p/2)}{\Gamma((p-1)/2)}\frac{1}{t^2} \\
&\leq \left(\frac{4}{\sqrt{p}}+\frac{4\sqrt{2}}{\pi\sqrt{p}}\frac{\Gamma(p/2)}{\Gamma((p-1)/2)}\right)\frac{\sqrt{p}}{t^2}.
\end{aligned}$$

Since

$$\lim_{k\to+\infty}\left(\frac{4}{\sqrt{k}}+\frac{4\sqrt{2}}{\pi\sqrt{k}}\frac{\Gamma(k/2)}{\Gamma((k-1)/2)}\right)$$

exists by Stirling's formula

$$\lim_{x\to+\infty}\frac{\Gamma(x+1)}{x^x\exp(-x)\sqrt{2\pi x}}=1,$$

we can find a c_2 such that

$$\frac{4}{\sqrt{k}}+\frac{4\sqrt{2}}{\pi\sqrt{k}}\frac{\Gamma(k/2)}{\Gamma((k-1)/2)}\leq c_2$$

for all k. Hence

$$f_{u_{pq}}(t)\leq c_2\sqrt{p}/t^2.$$

Therefore,

$$\begin{aligned}\text{Prob}(\|U\|_\infty > r) &\le \sum_{p=1}^{m}\sum_{q=p}^{m} P(|u_{pq}| > r/m)\\ &= \sum_{p=1}^{m}\sum_{q=p}^{m}\int_{|t|>r/m} f_{pq}(t)\,dt\\ &\le \sum_{p=1}^{m}\sum_{q=p}^{m}\int_{|t|>r/m}\frac{c_2\sqrt{p}}{t^2}dt\\ &\le \frac{c_2c_3}{r}m^{7/2}\end{aligned}$$

for some $c_3 > 0$, independent of m. The existence of c_3 is due to the existence of the limit

$$\lim_{k\to+\infty}\frac{1}{k^{5/2}}\sum_{p=1}^{k}(k-p+1)\sqrt{p} = \int_0^1 (1-t)\sqrt{t}\,dt.$$

We set $c_1 = c_2c_3$ and then (8) is proven. To prove the first inequality in the theorem, we note that the expected value μ and the variance σ^2 of the variable $x_1 \equiv \sum_{q=1}^{m}|x_{1q}|$ are

$$\mu = m\sqrt{\frac{2}{\pi}}, \qquad \sigma^2 = \left(1-\frac{2}{\pi}\right)m.$$

Setting $\varepsilon = m\sqrt{1-\frac{2}{\pi}}$ in Chebyshev's inequality [1, p. 183]

$$\text{Prob}\left(|x_1-\mu| \ge \varepsilon\right) \le \frac{\sigma^2}{\varepsilon^2},$$

we have

$$\text{Prob}\left(x_1 < mc_4\right) \le \frac{1}{m}, \tag{9}$$

where $c_4 = \sqrt{\frac{2}{\pi}} - \sqrt{1-\frac{2}{\pi}}$. Combining (8) and (9), we find

$$\begin{aligned}\text{Prob}\left(\rho_U > r\right) &= \text{Prob}\left(\|U\|_\infty > r\|A\|_\infty\right)\\ &\le \text{Prob}\left(\|U\|_\infty > rx_1\right)\\ &= \text{Prob}\left(\|U\|_\infty > rx_1,\, x_1 \ge mc_4\right) + \text{Prob}\left(\|U\|_\infty > rx_1,\, mc_4 > x_1 > 1\right)\\ &\quad + \text{Prob}\left(\|U\|_\infty > rx_1,\, x_1 \le 1\right)\\ &\le \text{Prob}\left(\|U\|_\infty > mrc_4\right) + \min\left(\text{Prob}\left(\|U\|_\infty > r\right),\, \text{Prob}\left(x_1 < mc_4\right)\right)\\ &\quad + \text{Prob}\left(|x_{1q}| \le 1\,\forall\, 1 \le q \le m\right)\\ &\le \frac{c_1}{c_4 r}m^{5/2} + \min\left(\frac{c_1}{r}m^{7/2},\, \frac{1}{m}\right) + \left(\frac{1}{\sqrt{2\pi}}\int_{-1}^{1}\exp\left(-\frac{1}{2}t^2\right)dt\right)^m.\end{aligned}$$

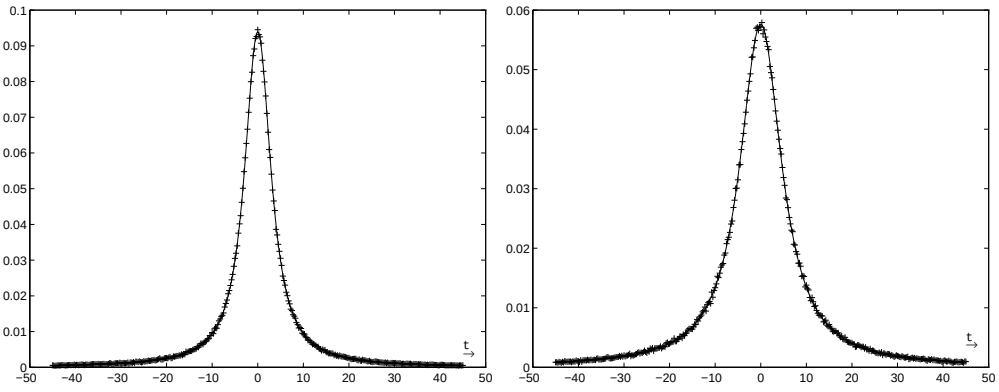
Finally,

$$\begin{aligned}\text{Prob}\,(\rho_L > r) &\leq \sum_{p=2}^{m}\sum_{q=1}^{p-1}\text{Prob}\left(|l_{pq}| > \frac{r-1}{m-1}\right)\\ &= \frac{1}{\pi}\sum_{p=2}^{m}\sum_{q=1}^{p-1}\int_{|t|>\frac{r-1}{m-1}}\frac{1}{1+t^2}dt\\ &\leq \frac{c_5}{r}m^3\end{aligned}$$

for some c_5. □

6. Numerical experiments. In this section, we present numerical results to support Theorems 1–4. All our calculations have been carried out in MATLAB 4.2c on SUN workstations.

In our first experiment, 595,000 matrices of dimension $m = 31$ were selected at random from the class $\mathcal{N}_{31}(O, I)$. Then GE was applied to each of the matrices and then statistics on the elements $l_{13,12}$, $l_{30,29}$, $u_{12,12}$, and $u_{31,31}$ were accumulated. The data are plotted in Figures 1 and 2 together with the corresponding functions indicated in Theorems 1 and 2. In order to make clearer the difference between Figures 1(a) and 1(b), we present them together in Figure 3(a).

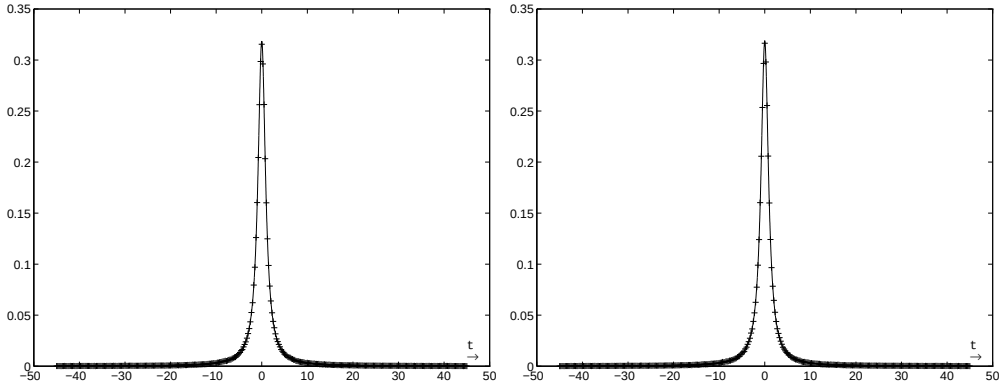


(a) (b)

FIG. 1. (a) *Distribution of* $u_{12,12}$: *observed* (+), *predicted* (-). (b) *Distribution of* $u_{31,31}$: *observed* (+), *predicted* (-).

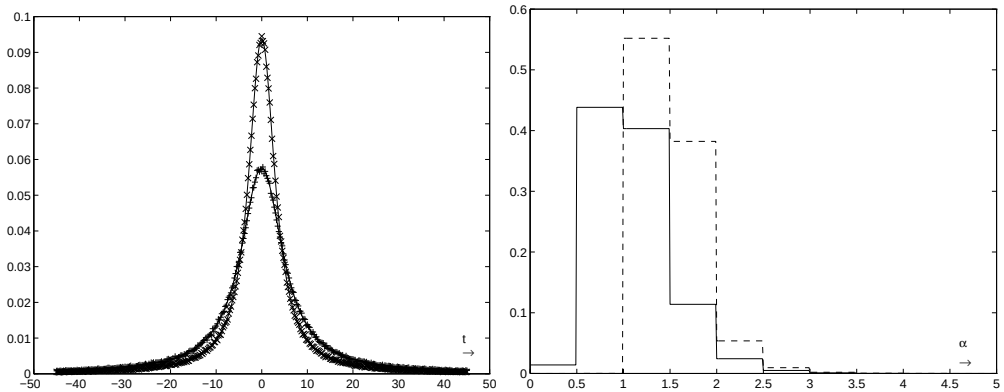
The purpose of our second experiment was to test formula (7). Matrices of several dimensions m were selected at random from $\mathcal{N}_m(O, I)$, with the sample size varying. A few tolerances ϵ were used. The results are outlined in Table 1. The frequency column of the table provides the numbers of matrices which, in their LU factors, have at least one u_{pp} less than ϵ in magnitude. By comparing with the empirical probabilities, we conclude that the bound given in (7) is a fairly tight one.

Finally, if we set $r = m^\alpha, \alpha > 2.5$ for ρ_U and $\alpha > 3$ for ρ_L, in Theorem 4, then we can see that the probabilities $\text{Prob}(\rho_L > m^\alpha)$ and $\text{Prob}(\rho_U > m^\alpha)$ decrease with m increasing. In fact, empirically this is true even for smaller α, say, $\alpha > 1.5$ for both ρ_L and ρ_U, as illustrated in Figures 3(b) and 4. In this experiment, we chose sample



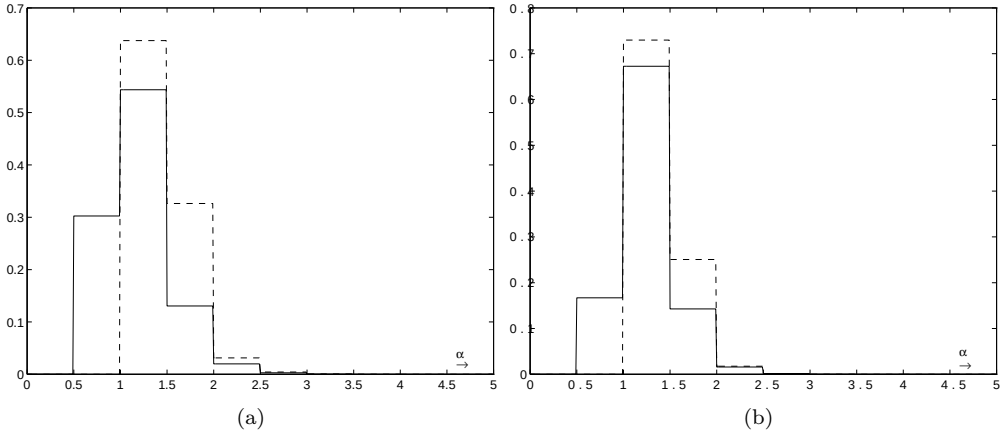
(a) (b)

FIG. 2. (a) *Distribution of* $l_{13,12}$: *observed* (+), *predicted* (-). (b) *Distribution of* $l_{30,29}$: *observed* (+), *predicted* (-).



(a) (b)

FIG. 3. (a) *Overlap of Figures* 1(a) *and* 1(b). (b) *Percentage frequency distributions of* ρ_L (*dashed*) *and* ρ_U (*solid*). $m = 25$.



(a) (b)

FIG. 4. *Percentage frequency distributions of* ρ_L (*dashed*) *and* ρ_U (*solid*). (a) $m = 50$. (b) $m = 100$.

TABLE 1
Probabilities of small pivot.

m	ϵ	*Sample size*	*Frequency*	*Empirical probability*	*Theoretical bound*
25	10^{-5}	10^5	5	5×10^{-5}	8.2853×10^{-5}
50	10^{-3}	10^4	90	0.009	0.012
50	10^{-4}	10^4	8	8×10^{-4}	0.0012
50	10^{-5}	10^4	0	0	1.2014×10^{-4}
50	10^{-5}	10^5	9	9×10^{-5}	1.2014×10^{-4}
75	10^{-3}	10^4	89	0.0089	0.0149
75	10^{-4}	10^4	8	8×10^{-4}	0.0015
75	10^{-5}	10^4	0	0	1.4876×10^{-4}
100	10^{-3}	10^4	115	0.0115	0.0173

sizes to be 968,500, 365,500, and 98,000 for $m = 25, 50$, and 100, respectively. In each sample, we calculated ρ_L and ρ_U for each matrix X. Then the data of ρ_L and ρ_U were grouped into ten classes, respectively. In the case of ρ_L, for example, the first class consists of matrices X with $m^0 \le \rho_L < m^{0.5}$, the second class with $m^{0.5} \le \rho_L < m^1$, the third one with $m^1 \le \rho_L < m^{1.5}$, and so on. The number of matrices in each class was then divided by the corresponding sample size to get the percentage frequency to the class. The distributions have been plotted in the form of histograms. Empirically, there is a tendency that $\text{Prob}(m \le \rho_L < m^{1.5})$ and $\text{Prob}(m \le \rho_U < m^{1.5})$ tend to one as m goes to infinity.

7. Appendix.

LEMMA 3.

$$\iiiint_{\Omega} w^{p-1} \exp\left(-\frac{1}{2}\left(x^2+y^2+z^2+w^2\right)\right) dx\, dy\, dz\, dw$$

$$= \int_{-\infty}^{\alpha} dt \int_{-\infty}^{\infty} dx \int_0^{\infty} \frac{w^p}{w^2+x^2} \exp\left(-\frac{1}{2}\left((x+t)^2+w^2\right)\right) dw,$$

where $\Omega = \{(x,y,z,w) \mid x - yw/z \le \alpha, w > 0, z > 0\}$ *and* $1 \le p$.

Proof.

$$F(\alpha) \equiv \iiiint_{\Omega} w^{p-1} \exp\left(-\frac{1}{2}\left(x^2+y^2+z^2+w^2\right)\right) dx\, dy\, dz\, dw$$

$$= \int_{-\infty}^{\infty} dx \int_0^{\infty} dw \int_0^{\infty} dz \int_{(x-\alpha)z/w}^{\infty} w^{p-1} \exp\left(-\frac{1}{2}\left(x^2+y^2+z^2+w^2\right)\right) dy.$$

Letting $y = uz$, we find

$$F(\alpha) = \int_{-\infty}^{\infty} dx \int_0^{\infty} dw \int_0^{\infty} dz \int_{(x-\alpha)/w}^{\infty} z w^{p-1} \exp\left(-\frac{1}{2}\left(x^2+u^2z^2+z^2+w^2\right)\right) du$$

$$= \int_{-\infty}^{\infty} dx \int_0^{\infty} dw \int_{(x-\alpha)/w}^{\infty} du \int_0^{\infty} z w^{p-1} \exp\left(-\frac{1}{2}\left(x^2+u^2z^2+z^2+w^2\right)\right) dz$$

$$= \int_{-\infty}^{\infty} dx \int_0^{\infty} dw \int_{(x-\alpha)/w}^{\infty} \frac{w^{p-1}}{1+u^2} \exp\left(-\frac{1}{2}(x^2+w^2)\right) du.$$

Letting $u = v/w$, this can then be written as

$$F(\alpha) = \int_{-\infty}^{\infty} dx \int_{0}^{\infty} dw \int_{x-\alpha}^{\infty} \frac{w^p}{w^2+v^2} \exp\left(-\frac{1}{2}(x^2+w^2)\right) dv$$
$$= \int_{-\infty}^{\infty} dx \int_{x-\alpha}^{\infty} dv \int_{0}^{\infty} \frac{w^p}{w^2+v^2} \exp\left(-\frac{1}{2}(x^2+w^2)\right) dw.$$

Finally, letting $v = x - t$, we have

$$F(\alpha) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\alpha} dt \int_{0}^{\infty} \frac{w^p}{w^2+(x-t)^2} \exp\left(-\frac{1}{2}(x^2+w^2)\right) dw$$
$$= \int_{-\infty}^{\alpha} dt \int_{-\infty}^{\infty} dx \int_{0}^{\infty} \frac{w^p}{w^2+(x-t)^2} \exp\left(-\frac{1}{2}(x^2+w^2)\right) dw$$
$$= \int_{-\infty}^{\alpha} dt \int_{-\infty}^{\infty} dx \int_{0}^{\infty} \frac{w^p}{w^2+x^2} \exp\left(-\frac{1}{2}\left((x+t)^2+w^2\right)\right) dw. \qquad \square$$

LEMMA 4.

$$\int_{0}^{\infty} dx \int_{-\infty}^{\infty} \frac{x^{p-1}}{x^2+y^2} \exp\left(-\frac{1}{2}\left(x^2+(y+t)^2\right)\right) dy$$
$$= 2^{(p-1)/2}\sqrt{\pi}\,\Gamma\left(\frac{p}{2}\right)\left(\sum_{i=0}^{\lfloor\frac{p-3}{2}\rfloor} \xi_{i,p}\, t^{-2i-2} + (-1)^{\lfloor (p-1)/2\rfloor}\,\zeta_p\, t^{-p+1} \exp\left(-\frac{1}{2}t^2\right)\,\phi_p(t)\right),$$

where

$$\xi_{i,p} = \begin{cases} (-1)^i \prod_{j=0}^{i-1}(p-2j-3), & i > 0, \\ 1, & i = 0, \end{cases}$$

$$\zeta_p = \begin{cases} (p-3)!!, & p > 3, \\ 1, & p = 2,\ 3, \end{cases}$$

$$\phi_p(t) = \left(\int_0^t \exp\left(\frac{1}{2}x^2\right) dx\right)^{p-1-2\lfloor (p-1)/2\rfloor},$$

and where $-\infty < t < \infty$, $2 \le p$.

Proof.

$$
\begin{aligned}
f(t) &\equiv \int_0^\infty dx \int_{-\infty}^\infty \frac{x^{p-1}}{x^2+y^2} \exp\left(-\frac{1}{2}\left(x^2+(y+t)^2\right)\right) dy \\
&= \exp\left(-\frac{1}{2}t^2\right) \int_0^\infty dx \int_{-\infty}^\infty \frac{x^{p-1}}{x^2+y^2} \exp\left(-\frac{1}{2}\left(x^2+y^2\right)\right) \exp\left(-yt\right) dy \\
&= \exp\left(-\frac{1}{2}t^2\right) \int_0^\infty dx \int_{-\infty}^\infty \frac{x^{p-1}}{x^2+y^2} \exp\left(-\frac{1}{2}\left(x^2+y^2\right)\right) \sum_{n=0}^\infty \frac{(-yt)^n}{n!} dy \\
&= \exp\left(-\frac{1}{2}t^2\right) \sum_{n=0}^\infty \frac{(-t)^n}{n!} \int_0^\infty dx \int_{-\infty}^\infty \frac{x^{p-1}y^n}{x^2+y^2} \exp\left(-\frac{1}{2}\left(x^2+y^2\right)\right) dy \\
&= 2\,\exp\left(-\frac{1}{2}t^2\right) \sum_{n=0}^\infty \frac{t^{2n}}{(2n)!} \int_0^\infty dx \int_0^\infty \frac{x^{p-1}y^{2n}}{x^2+y^2} \exp\left(-\frac{1}{2}\left(x^2+y^2\right)\right) dy \\
&= 2\,\exp\left(-\frac{1}{2}t^2\right) \sum_{n=0}^\infty \frac{t^{2n}}{(2n)!} \int_0^\infty dz \int_0^\infty \frac{z^{2n}}{1+z^2}\, x^{2n+p-2} \exp\left(-\frac{1}{2}x^2\left(1+z^2\right)\right) dx,
\end{aligned}
$$

where $y = xz$. Let $w = x^2\left(1+z^2\right)/2$. Then, with $\mathbf{B}(m,n)$ denoting the beta function, we have

$$
\begin{aligned}
f(t) &= 2^{(p-1)/2} \exp\left(-\frac{1}{2}t^2\right) \sum_{n=0}^\infty \frac{2^n t^{2n}}{(2n)!} \\
&\quad \times \int_0^\infty dz \int_0^\infty \frac{z^{2n}}{(1+z^2)^{n+(p+1)/2}}\, w^{n+(p-3)/2} \exp\left(-w\right) dw \\
&= 2^{(p-3)/2} \exp\left(-\frac{1}{2}t^2\right) \sum_{n=0}^\infty \frac{2^n t^{2n}}{(2n)!} \Gamma\left(n+\frac{p-1}{2}\right) \mathbf{B}\left(\frac{p}{2}, n+\frac{1}{2}\right) \\
&= 2^{(p-3)/2} \exp\left(-\frac{1}{2}t^2\right) \sum_{n=0}^\infty \frac{2^n t^{2n}}{(2n)!} \Gamma\left(n+\frac{p-1}{2}\right) \frac{\Gamma\left(\frac{p}{2}\right)\Gamma\left(n+\frac{1}{2}\right)}{\Gamma\left(n+\frac{p+1}{2}\right)} \\
&= 2^{(p-1)/2}\sqrt{\pi}\,\Gamma\left(\frac{p}{2}\right) \exp\left(-\frac{1}{2}t^2\right) \sum_{n=0}^\infty \frac{1}{2n+p-1} \frac{(2n-1)!!}{(2n)!}\, t^{2n} \\
&= 2^{(p-1)/2}\sqrt{\pi}\,\Gamma\left(\frac{p}{2}\right) \exp\left(-\frac{1}{2}t^2\right) \sum_{n=0}^\infty \frac{1}{n!\,2^n} \frac{1}{2n+p-1}\, t^{2n} \\
&= 2^{(p-1)/2}\sqrt{\pi}\,\Gamma\left(\frac{p}{2}\right) \exp\left(-\frac{1}{2}t^2\right) t^{-p+1} \sum_{n=0}^\infty \frac{1}{n!\,2^n} \frac{1}{2n+p-1}\, t^{2n+p-1} \\
&\equiv 2^{(p-1)/2}\sqrt{\pi}\,\Gamma\left(\frac{p}{2}\right) \exp\left(-\frac{1}{2}t^2\right) t^{-p+1} g(t),
\end{aligned}
$$

where here and below we define $0! = 0!! = (-1)!! = 1$. Since

$$
\frac{d}{dt}g(t) = \sum_{n=0}^\infty \frac{1}{n!\,2^n} t^{2n+p-2} = t^{p-2} \exp\left(\frac{1}{2}t^2\right)
$$

we find

$$\begin{aligned} g(t) &= \int_0^t x^{p-2} \exp\left(\frac{1}{2}x^2\right) dx \\ &= \exp\left(\frac{1}{2}t^2\right) \sum_{i=0}^{\lfloor \frac{p-3}{2} \rfloor} (-1)^i \prod_{j=0}^{i-1} (p-2j-3)\, t^{p-2i-3} \\ &\quad + (-1)^{\lfloor (p-1)/2 \rfloor} (p-3)!! \left(\int_0^t \exp\left(\frac{1}{2}x^2\right) dx \right)^{p-1-2\lfloor (p-1)/2 \rfloor} \end{aligned}$$

by integration by parts, and then the desired result follows. □

LEMMA 5.

$$\begin{aligned} &\int \cdots \int_\Omega x_1^{q-2} \exp\left(-\frac{1}{2}\sum_{i=1}^6 x_i^2\right) \prod_{i=1}^6 dx_i \\ &\quad = \pi^{1/2} 2^{(q-2)/2} \Gamma((q+1)/2) \int_{-\infty}^{\alpha} dy_1 \int_{-\infty}^{\infty} dy_2 \int_{-\infty}^{\infty} dy_3 \int_{-\infty}^{\infty} \frac{|y_3|}{(1+y_2^2+y_4^2)^{(q+1)/2}} \\ &\qquad \times \frac{1}{\left(1+(y_1y_3+y_2)^2+(y_3+y_4)^2\right)^{3/2}} dy_4, \end{aligned}$$

where $q \geq 2$ *and* $\Omega = \{(x_1, \ldots, x_6) \mid (x_1x_3 - x_2x_4)/(x_1x_5 - x_2x_6) \leq \alpha,\ x_1 > 0,\ x_2 > 0\}$.

Proof. Let

$$x_3 = (y_1 + y_2)x_2, \qquad x_4 = x_1y_2,$$

$$x_5 = (y_3 + y_4)x_2, \qquad x_6 = x_1y_4.$$

Then

$$\begin{aligned} F(\alpha) &\equiv \int \cdots \int_\Omega x_1^{q-2} \exp\left(-\frac{1}{2}\sum_{i=1}^6 x_i^2\right) \prod_{i=1}^6 dx_i \\ &= \int \cdots \int_{y_1/y_3 \leq \alpha} x_1^q x_2^2 \exp\left(-\frac{1}{2}\left(x_1^2\left(1+y_2^2+y_4^2\right) + x_2^2\left(1+(y_1+y_2)^2+(y_3+y_4)^2\right)\right)\right) \\ &\qquad \times dx_1\, dx_2 \prod_{i=1}^4 dy_i \\ &= \iiiint_{y_1/y_3 \leq \alpha} \left(\int_0^\infty x_1^q \exp\left(-\frac{1}{2}x_1^2(1+y_2^2+y_4^2)\right) dx_1 \right) \\ &\qquad \times \left(\int_0^\infty x_2^2 \exp\left(-\frac{1}{2}x_2^2\left(1+(y_1+y_2)^2+(y_3+y_4)^2\right)\right) dx_2 \right) \prod_{i=1}^4 dy_i \\ &= \pi^{1/2} 2^{(q-2)/2} \Gamma((q+1)/2) \iiiint_{y_1/y_3 \leq \alpha} \frac{1}{\left(1+(y_1+y_2)^2+(y_3+y_4)^2\right)^{3/2}} \\ &\qquad \times \frac{1}{(1+y_2^2+y_4^2)^{(q+1)/2}} \prod_{i=1}^4 dy_i \end{aligned}$$

$$= \pi^{1/2}2^{(q-2)/2}\Gamma((q+1)/2)\iint_{y_1/y_3\le\alpha} dy_1\,dy_3\int_{-\infty}^{\infty}dy_2\int_{-\infty}^{\infty}\frac{1}{(1+y_2^2+y_4^2)^{(q+1)/2}}$$
$$\times\frac{1}{\left(1+(y_1+y_2)^2+(y_3+y_4)^2\right)^{3/2}}\,dy_4.$$

Since

$$\iint_{x/y\le\alpha} f(x,y)\,dx\,dy = \int_{-\infty}^{\alpha}dx\int_{-\infty}^{\infty}f(xy,y)|y|\,dy$$

we have

$$F(\alpha) = \pi^{1/2}2^{(q-2)/2}\Gamma((q+1)/2)\int_{-\infty}^{\alpha}dy_1\int_{-\infty}^{\infty}dy_3\int_{-\infty}^{\infty}dy_2\int_{-\infty}^{\infty}\frac{|y_3|}{(1+y_2^2+y_4^2)^{(q+1)/2}}$$
$$\times\frac{1}{\left(1+(y_1y_3+y_2)^2+(y_3+y_4)^2\right)^{3/2}}dy_4. \qquad \square$$

LEMMA 6. *Let*

$$f(t) = \frac{q-1}{4\pi^2}\int_{-\infty}^{\infty}dx_1\int_{-\infty}^{\infty}dx_2\int_{-\infty}^{\infty}\frac{|x_1|}{\left(1+(x_2+x_1t)^2+(x_3+x_1)^2\right)^{3/2}}$$
$$\times\frac{1}{(1+x_2^2+x_3^2)^{(q+1)/2}}\,dx_3,$$

where $2\le q$ *and* $-\infty<t<\infty$. *Then*

$$f(t) = \frac{1}{\pi}\frac{1}{1+t^2}.$$

Proof. We rewrite the expression of $f(t)$ as

$$f(t) = \frac{q-1}{4\pi^2}\frac{1}{1+t^2}\int_{-\infty}^{\infty}dx_1\int_{-\infty}^{\infty}dx_2\int_{-\infty}^{\infty}\frac{|x_1|}{\left(1+(x_2+cx_1)^2+(x_3+sx_1)^2\right)^{3/2}}$$
$$\times\frac{1}{(1+x_2^2+x_3^2)^{(q+1)/2}}\,dx_3,$$

where $c = t/\sqrt{1+t^2}$ and $s = 1/\sqrt{1+t^2}$. Let

$$x_1 = y_1, \qquad x_2 = y_1(cy_2 - sy_3), \qquad x_3 = y_1(sy_2+cy_3).$$

Then

$$f(t) = \frac{q-1}{4\pi^2}\frac{1}{1+t^2}\int_{-\infty}^{\infty}dy_1\int_{-\infty}^{\infty}dy_2\int_{-\infty}^{\infty}\frac{y_1^2|y_1|}{\left(1+y_1^2\left((1+y_2)^2+y_3^2\right)\right)^{3/2}}$$
$$\times\frac{1}{(1+y_1^2\,(y_2^2+y_3^2))^{(q+1)/2}}\,dy_3$$
$$\equiv \frac{1}{\pi}\frac{1}{1+t^2}\xi.$$

Since

$$\int_{-\infty}^{\infty} f(t)dt = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{1}{1+t^2}dt = 1,$$

we have $\xi = 1$, and therefore the lemma follows. □

Acknowledgments. We wish to thank Professor Alan Edelman for providing us with several important references and suggestions on an earlier draft which improved the derivation of equations (1) and (4). In addition, we would like to thank Professors Gene Golub and Nick Trefethen for their valuable comments and suggestions on the earlier draft.

REFERENCES

[1] H. Cramér, *Mathematical Methods of Statistics,* Princeton University Press, 1946.
[2] A. Edelman, *Eigenvalues and condition numbers of random matrices,* SIAM J. Matrix Anal. Appl., 9 (1988), pp. 543–560.
[3] A. Edelman, *The distribution and moments of the smallest eigenvalue of a random matrix of Wishart type,* Linear Algebra Appl., 159 (1991), pp. 55–80.
[4] A. Edelman and W. Mascarenhas, *On the complete pivoting conjecture for a Hadamard matrix of order* 12, J. Linear and Multilinear Algebra, 38 (1995), pp. 181–187.
[5] L. V. Foster, *The probability of large diagonal elements in the QR factorization,* SIAM J. Sci. Stat. Comput., 11 (1990), pp. 531–544.
[6] L. V. Foster, *Gaussian elimination with partial pivoting can fail in practice,* SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1354–1362.
[7] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[8] D. S. Parker, *Random Butterfly Transformations with Applications in Computational Linear Algebra*, Tech. report CSD-950023, Computer Science Department, UCLA, 1995.
[9] D. S. Parker and B. Pierce, *The Randomizing FFT: An Alternative to Pivoting in Gaussian Elimination*, Tech. report CSD-950037, Computer Science Department, UCLA, 1995.
[10] J. W. Silverstein, *The smallest eigenvalue of a large-dimensional Wishart matrix,* Ann. Prob., 13 (1985), pp. 1364–1368.
[11] L. N. Trefethen and R. S. Schreiber, *Average-case stability of Gaussian elimination,* SIAM J. Matrix Anal. Appl., 11 (1990), pp. 335–360.
[12] H. F. Trotter, *Eigenvalue distributions of large Hermitian matrices*; *Wigner's semi-circle law and a theorem of Kac, Murdock, and Szegő,* Advances in Math., 54 (1984), pp. 67–82.
[13] A. M. Turing, *Rounding-off errors in matrix processes*, Quart. J. Mech. Appl. Math., 1 (1948), pp. 287–308.
[14] J. von Neumann and H. H. Goldstine, *Numerical inverting of matrices of high order*, Bull. Amer. Math. Soc., 53 (1947), pp. 1021–1099; in von Neumann's Collected Works, Vol. 5, A. H. Taub, ed., Pergamon, Elmsford, NY, 1963.
[15] J. von Neumann and H. H. Goldstine, *Numerical inverting of matrices of high order, Part* II, Proc. Amer. Math. Soc., 2 (1951), pp. 188-202; in von Neumann's Collected Works, Vol. 5, A. H. Taub, ed., Pergamon, Elmsford, NY, 1963.
[16] J. H. Wilkinson, *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., 8 (1961), pp. 281–330.
[17] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 2, p. 518, April 1997

018

COMMENTS ON NORMAL TOEPLITZ MATRICES BY FARENICK ET AL.*

KHAKIM D. IKRAMOV†

PII. S0895479897314796

In [1], the authors refer to [2] as the paper that gave an incentive to their work. Here we point out that the problem considered in [1] was also solved in the Russian literature. This was not known to the authors and might be of interest to readers of this journal. In the introduction of [1], the authors say,

In [2] Ikramov has shown that a normal Toeplitz matrix (of order at most 4) over the real field must be of one of four types: symmetric Toeplitz, skew-symmetric Toeplitz (up to the principal diagonal), circulant, or skew-circulant.

It would be correct to add that the main contribution of [2] was to prove, for arbitrary n, the necessity of some equalities for the entries of a normal Toeplitz matrix of order n. These equalities are in fact equivalent to the authors' equations (1) in the real case.

The full solution of the problem (of describing normal Toeplitz matrices) for the real case was given in [3]. The same year another and very elegant solution of the real problem was published in [4]. The paper [5] in the February 1996 issue of *Computational Mathematics and Mathematical Physics* solves the complex version of the problem, i.e., the version the authors of [1] are dealing with. A different and, again, quite ingenious proof is to appear this year in the November issue of the journal mentioned above [6].

REFERENCES

[1] D. R. FARENICK, M. KRUPNIK, N. KRUPNIK, AND W. Y. LEE, *Normal Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 1037–1043.

[2] KH. D. IKRAMOV, *Describing normal Toeplitz matrices*, Zh. Vychisl. Mat. i Mat. Fiz., 34 (1994), pp. 473–479 (in Russian). Comput. Math. Math. Phys., 34 (1994), pp. 399–404.

[3] KH. D. IKRAMOV, *Classification of normal Toeplitz matrices with real entries*, Mat. Zametki, 57 (1995), pp. 670–680 (in Russian). Math. Notes, 57 (1995), pp. 463–469.

[4] V. I. GEL'FGAT, *A normality criterium for Toeplitz matrices*, Zh. Vychisl. Mat. i Mat. Fiz., 35 (1995), pp. 1428–1432 (in Russian). Comput. Math. Math. Phys., 35 (1995), pp. 1147–1150.

[5] KH. D. IKRAMOV AND V. N. CHUGUNOV, *Normality condition for a complex Toeplitz matrix*, Zh. Vychisl. Mat. i Mat. Fiz., 36 (1996), pp. 3–10 (in Russian). Comput. Math. Math. Phys., 36 (1996), pp. 131–137.

[6] V. I. GEL'FGAT, *Commutativity conditions for Toeplitz matrices*, Zh. Vychisl. Mat. i Mat. Fiz., to appear.

* Received by the editors January 9, 1997; accepted for publication (in revised form) by P. Van Dooren January 13, 1997.
http://www.siam.org/journals/simax/18-2/31479.html

† Faculty of Computational Mathematics and Cybernetics, Moscow State University, 119899 Moscow, Russia (ikramov@cmc.msk.su).

List of Referees

The following people have reviewed papers for the SIAM Journal on Matrix Analysis and Applications from January 1, 1996 *to December* 31, 1996. *SIAM appreciates the efforts of these referees and gratefully acknowledges their contribution to the journal.*

D. Alpay
G. S. Ammar
M. Arioli
C. Ashcraft
O. Axelsson
J. Baglama
Z. Bai
Z.-J. Bai
R. B. Bapat
J. L. Barlow
W. Barrett
A. Barrlund
A. Ben-Israel
J. Berg
M. Berry
T. Bhattacharyya
R. Bini
A. Bjorck
A. Boettcher
A. W. Bojanczyk
D. Boley
K. N. Boyadzhiev
C. Brezinski
M. Bruehl
A. Bunse-Gerstner
R. Byers
X.-C. Cai
D. Calvetti
S. Campbell
S. Chandrasekaran
X.-W. Chang
F. Chatin-Chatelin
C.-M. Cheng
E. Chu
D. Clements
M. Cosnard
R. Cottle
K. Dackland
K. R. Davidson
T. Davis
J. Day
T. Dayar
J. Dehaene
I. de Hoyos
M. Deistler
J. Demmel
B. L. R. De Moor
P. Deuflhard
A. Dubrulle
I. Duff
A. Edelman
S. C. Eisenstat
L. El Ghaoui
E. Elmroth
L. Elsner
R. Entringer
O. Ernst
D. Farenick
H. Fassbender
B. Fischer
P. F. Fischer
D. Fokkema
A. Forsgren
L. Foster
V. Fraysse
G. M. Freimann
R. W. Freund
S. R. Gadre
P. Gahinet
K. Georg
A. George
A. Gerasoulis
J. Gilbert
L. Gleser
W. Grassman
E. Grimme
J. Gross
M. Gu
M. Gulliksson
M. Gutknecht
J.-P. A. Haeberly
S. Hammarling
M. Hanke
P. C. Hansen
B. Hanzon
T. Hara
C. He
G. Heinig
J. Hench
D. P. Heyman
D. J. Higham
N. J. Higham
D. Hinrichsen
M. Hochbruck
R. Horn
T. Huckle
I. Ipsen
T. Ito
F. Jarre
E. Jessup
J. Ji
X. Q. Jin
W. D. Joubert
B. Kalantari
L. Kaufman
C. Kenney
D. E. Keyes
D. Kincaid
F. Kittaneh
L. Knizhnerman
L. Kolotilina
A. Kovacec
A. Laratta
M. Laurent
R. Lehoucq
P. Lemmerling
J. Lewis
C.-K. Li
R.-C. Li
T.-Y. Li
S. Liu
I. Marek
R. Mathias
K. Meerbergen
C. Mehl
V. Mehrmann
R. Merris
V. Metz
C. Meyer
G. Miminis
P. Misra
R. Morgan
J. Moro
R. Nabben
N. M. Nachtigal
M. Nakamura
S. K. Narayan
L. Nazareth
H. Neudecker
M. Neumann
E. G. Ng
C. Oara
K. O'Brien
C. O'Cinneide
K. Okubo
D. Olesky
V. Olshevsky

C. Paige
H. Park
B. Parlett
R. V. Patel
M. D. Perlman
P. Petkov
B. Peyton
R. Plemmons
A. Pothen
K. M. Prasad
F. Pukelsheim
E. Quintana
V. Ramaswami
J. Reid
R. Renaut
S. Roch
C. Rodman
J. Rohn
C. Roos
U. Rothblum
A. Ruhe
Y. Saad
M. A. Saunders
A. H. Sayed
R. Schnabel
R. Schreiber
B. Shader
A. Shapiro
H. Shapiro
A. Sidi
G. Sleijpen
D. Stanford
G. Starke
M. Stewart
P. Stoica
G. P. H. Styan
J.-g. Sun
J. Swevers
V. Syrmos
D. B. Szyld
K. Takahashi
W. P. Tang
P. Tarazaga
D. Temme
C. Tomei
C. Tong
G. Trapp
M. Trummer
M. Tsatsomeros
M. Tůma
E. Tyrtyshnikov
F. Uhlig
L. Vandenberghe
H. A. van der Vorst
P. Van Dooren
S. Van Huffel
C. Van Loan
S. Vavasis
K. Veselic
U. Von Matt
P. Vu
B. Waldén
H. F. Walker
G. Wang
G. Wasilkowski
D. S. Watkins
L. Watson
J. R. Weaver
M. Weiss
H. J. Werner
H. K. Wimmer
F. Wirth
H. Woerdeman
H. Wolkowicz
P. Wortelboer
H. Wosniakowski
M. Wright
S. Wright
S.-P. Wu
Q. Ye
P. Zagalak
Z. Zeng
H. Zha

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 3, pp. 521–534, July 1997

001

THE MINIMUM EIGENVALUE OF A SYMMETRIC POSITIVE-DEFINITE TOEPLITZ MATRIX AND RATIONAL HERMITIAN INTERPOLATION*

WOLFGANG MACKENS† AND HEINRICH VOSS†

Abstract. A novel method for computing the minimal eigenvalue of a symmetric positive-definite Toeplitz matrix is presented. Similar to the algorithm of Cybenko and Van Loan, it is a combination of bisection and a root finding method. Both phases of the method are accelerated considerably by rational Hermitian interpolation of the secular equation. For randomly generated test problems of dimension 800 the average number of linear systems which must be solved to determine the smallest eigenvalue is 6.6, which reduces the computational cost of the method of Cybenko and Van Loan to approximately 35%. The method includes a rigorous error bound.

Key words. Toeplitz matrix, eigenvalue problem, rational Hermitian interpolation

AMS subject classification. 65F15

PII. S0895479895288851

1. Introduction.

In this paper we consider a method for computing the smallest eigenvalue λ_1 of a symmetric and positive-definite Toeplitz matrix T. This problem is of considerable interest in signal processing. Given the covariance sequence of the observed data, Pisarenko [13] suggested a method which determines the sinusoidal frequencies from the eigenvector of the covariance matrix associated with the minimum eigenvalue of T.

Cybenko and Van Loan [3] introduced an algorithm which takes advantage of the Levinson–Durbin method for shifted matrices combining a bisection method and Newton's method for the secular equation. Hu and Kung [9] considered a safeguarded inverse iteration with shifts and Huckle [10], [11] studied the spectral transformation Lanczos method. Trench [14] and Noor and Morgera [12] generalized the method of Cybenko and Van Loan to the computation of the complete spectrum.

Our method generalizes the approach of Cybenko and Van Loan. The form of the secular equation suggests basing a root finding procedure on rational Hermitian interpolation. Similar to Newton's method, the originating algorithm converges monotonely decreasing and quadratically to λ_1. The local convergence is guaranteed to be faster than Newton's method and its global behavior is much better. Moreover, the bisection phase of the procedure can be accelerated by rational Hermitian interpolation.

Our paper is organized as follows. In section 2 we briefly sketch the method of Cybenko and Van Loan. Section 3 describes the connection of the secular equation to condensation methods. It introduces the rational Hermitian interpolation, which is the basis of various enhancements of the algorithm of Cybenko and Van Loan. Moreover, it contains a lower bound of the smallest eigenvalue based on Hermitian quadratic interpolation of the secular equation. In section 4 we give a MATLAB program of the originating procedure and in section 5 we discuss its numerical behavior.

*Received by the editors July 7, 1995; accepted for publication by G. Cybenko June 14, 1996.
http://www.siam.org/journals/simax/18-3/28885.html
†Hamburg University of Technology, Arbeitsbereich Mathematik, Kasernenstrasse 12, 2017H3 Hamburg, Federal Republic of Germany (mackens@tu-harburg.de, voss@tu-harburg.de).

2. The method of Cybenko and Van Loan. In this section we briefly sketch the approach of Cybenko and Van Loan to the computation of the smallest eigenvalue of a real symmetric positive-definite Toeplitz matrix and discuss two improvements of that method.

Let $T \in \mathbb{R}^{(n,n)}$ be a symmetric positive-definite Toeplitz matrix. Without restriction of generality we assume that its diagonal is normalized, and we consider the following partition:

$$T = \begin{pmatrix} 1 & t^T \\ t & G \end{pmatrix}.$$

Then it is well known that the eigenvalues of T and G are real and positive and satisfy an interlacing property

$$\lambda_1 \le \omega_1 \le \lambda_2 \le \cdots \le \omega_{n-1} \le \lambda_n,$$

where λ_j and ω_j are the jth smallest eigenvalues of T and G, respectively.

Eliminating the variables $x_2, \ldots, x_n$ from the system of equations

$$\begin{pmatrix} 1-\lambda & t^T \\ t & G-\lambda I \end{pmatrix} x = 0$$

that characterizes the eigenvalues of T, one obtains

$$\bigl(1 - \lambda - t^T(G-\lambda I)^{-1}t\bigr)x_1 = 0.$$

If an eigenvalue λ_j of T is not an eigenvalue of G and x is a corresponding eigenvector, then its first component x_1 is different from zero. Hence, λ_j is a root of the secular equation

$$f(\lambda) := -1 + \lambda + t^T(G-\lambda I)^{-1}t = 0. \tag{2.1}$$

We assume that

$$\lambda_1 < \omega_1.$$

Then λ_1 is the smallest root of f. In the interval $(0, \omega_1)$ it holds that

$$f'(\lambda) = 1 + \|(G-\lambda I)^{-1}t\|_2^2 > 1, \tag{2.2}$$

$$f''(\lambda) = 2t^T(G-\lambda I)^{-3}t > 0. \tag{2.3}$$

Therefore, f is strictly monotonely increasing and strictly convex there, and it is well known that for every initial value $\mu_0 \in (\lambda_1, \omega_1)$ Newton's method converges monotonely decreasing and quadratically to λ_1.

Equations (2.1) and (2.2) show that a Newton step can be performed in the following way:

$$\begin{aligned} &\text{Solve } (G-\mu_k I)w = -t \text{ for } w \\ &\text{and set } \mu_{k+1} := \mu_k - \frac{-1+\mu_k - w^T t}{1+\|w\|_2^2}, \end{aligned}$$

where the Yule–Walker system

$$(G-\mu I)w = -t \tag{2.4}$$

can be solved by Durbin's algorithm (cf. [7, p. 184 ff]) requiring $2n^2$ flops.

An initial value μ_0 for Newton's method can be obtained by a bisection process. If μ is not in the spectrum of any of the principal submatrices of $T-\mu I$ then Durbin's algorithm applied to $(T-\mu I)/(1-\mu)$ determines a lower triangular matrix

$$L = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \ell_{21} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ \ell_{n1} & \ell_{n2} & \dots & 1 \end{pmatrix}$$

such that

$$\frac{1}{1-\mu} L(T-\mu I)L^T = D := \operatorname{diag}\{1, E_1, \dots, E_{n-1}\}. \tag{2.5}$$

If $\tilde{L}$ is obtained from L by dropping the last row and last column then obviously

$$\frac{1}{1-\mu} \tilde{L}(G-\mu I)\tilde{L}^T = \tilde{D} := \operatorname{diag}\{1, E_1, \dots, E_{n-2}\}.$$

Hence, Sylvester's law of inertia yields the following method to determine the position of a parameter $\mu \in (0,1)$:

(i) if $E_j > 0$ for $j = 1, \dots, n-1$, then $\mu < \lambda_1$,
(ii) if $E_j > 0$ for $j = 1, \dots, n-2$ and $E_{n-1} \le 0$, then $\mu \in [\lambda_1, \omega_1)$,
(iii) if $E_j < 0$ for some $j \in \{1, \dots, n-2\}$, then $\mu > \omega_1$.

For the determination of an upper bound of λ_1 to start the bisection process Cybenko and Van Loan suggested four methods. They recommended using the following one. Determine $\varepsilon > 0$ such that

$$\tilde{T} := T + \varepsilon(e^1(e^n)^T + e^n(e^1)^T)$$

is singular and positive semidefinite, where e^1 and e^n are the unit vectors containing a 1 in their first and last components, respectively. Then it follows from Rayleigh's principle that $\lambda_1 \le \varepsilon$:

$$\begin{aligned} 0 &= \min_{x \neq 0} \frac{1}{\|x\|_2^2} \Big(x^T T x + \varepsilon x^T (e^1(e^n)^T + e^n(e^1)^T) x \Big) \\ &= \min_{x \neq 0} \frac{1}{\|x\|_2^2} \Big(x^T T x + 2\varepsilon x_1 x_n \Big) \\ &\ge \min_{x \neq 0} \frac{x^T T x}{\|x\|_2^2} + 2 \min_{x \neq 0} \frac{\varepsilon x_1 x_n}{\|x\|_2^2} = \lambda_1 - \varepsilon. \end{aligned}$$

ε can easily be determined from the data in Durbin's algorithm for the system $Gw = -t$.

At nearly the same cost one can get a much better bound. Let $w := -G^{-1}t$ be the solution of the Yule–Walker system. Then

$$y := \frac{1}{1 + t^T w} \begin{pmatrix} 1 \\ w \end{pmatrix} = T^{-1} e^1$$

is the first iterate of the inverse iteration with shift parameter 0 starting with the unit vector e^1 which can be expected to be a decent approximation of the eigenvector corresponding to the smallest eigenvalue λ_1. The Rayleigh quotient

$$R(y) := \frac{y^T T y}{y^T y} = \frac{1 + t^T w}{1 + \|w\|_2^2} \tag{2.6}$$

TABLE 1
Eigenvalues, subeigenvalues, and initial guesses, $n = 800$.

Expl.	λ_1	ω_1	$R(y)$	ε
1	2.652607 $E-5$	2.735377 $E-5$	2.90 $E-4$	1.73 $E-1$
2	3.477361 $E-6$	3.479143 $E-6$	7.01 $E-5$	1.44 $E-1$
3	3.314204 $E-5$	3.466962 $E-5$	2.54 $E-4$	1.57 $E-1$
4	4.349535 $E-5$	4.382751 $E-4$	4.48 $E-4$	1.29 $E-1$
5	2.615966 $E-4$	2.712415 $E-4$	4.34 $E-3$	1.90 $E-1$
6	2.918829 $E-5$	3.045659 $E-5$	6.06 $E-4$	1.84 $E-1$
7	1.976995 $E-6$	2.029093 $E-6$	2.30 $E-5$	1.03 $E-1$
8	8.864895 $E-4$	9.179502 $E-4$	6.95 $E-3$	1.36 $E-1$
9	2.377467 $E-6$	2.485892 $E-6$	3.84 $E-5$	1.17 $E-1$
10	1.485040 $E-4$	1.533171 $E-4$	2.38 $E-3$	1.66 $E-1$

is an upper bound of λ_1 which should not be too bad. Bound (2.6) was already given by Dembo [5]. Using the power series representation of the secular equation, he also proved a tighter bound. However, to evaluate this one we would have to solve the linear system $Gz = w$ for z. Using Levinson's algorithm, this requires $2n^2$ additional flops, i.e., the cost of one step of the bisection process. This normally does not pay.

We have not been able to prove that $R(y)$ is always a smaller bound than ε. However, in all the examples that we considered it was better by orders of magnitude.

Example 1. Table 1 contains λ_1, ω_1 as well as the bounds $R(y)$ and ε for ten examples of dimension $n = 800$ of the form

$$T = m \sum_{k=1}^{n} \eta_k T_{2\pi\theta_k}, \tag{2.7}$$

where m is chosen such that T has a normalized diagonal,

$$T_\theta = (t_{ij}) = (\cos(\theta(i-j))),$$

and η_k and θ_k are uniformly distributed random numbers taken from $[0, 1]$ (cf. Cybenko and Van Loan [3]).

A second improvement of the approach of Cybenko and Van Loan is at hand. If in the bisection process a parameter $\mu \in (0, \lambda_1)$ is tested then the only advantage that is taken from the Durbin algorithm is the information that μ is in $(0, \lambda_1)$. The vector $w := -(G - \mu I)^{-1}t$ is not used to enhance the interval that contains λ_1. However, with two additional inner products one can evaluate the Newton iterate

$$\tilde{\mu} := \frac{1 + t^T w + \mu \|w\|_2^2}{1 + \|w\|_2^2}$$

for $f(\lambda) = 0$ with initial guess μ. By the monotonicity and convexity of f this iterate is an upper bound of λ_1, and it should be chosen as the new upper bound in the bisection process if it is smaller than the current upper bound.

Notice that $\tilde{\mu}$ can be expected to be a good approximation of λ_1 since it can be interpreted as the Rayleigh quotient of T at $y := \begin{pmatrix} 1 \\ w \end{pmatrix}$. From

$$(G - \mu I)w = -t$$

TABLE 2
Number of flops for 100 *test examples.*

dimension	Cybenko –Van Loan	improved initial upper bound	improved subsequent upper bounds
50	6.37 *E*6	5.19 *E*6 (81.6%)	4.55 *E*6 (71.4%)
100	2.71 *E*7	2.23 *E*7 (82.1%)	1.92 *E*7 (70.9%)
200	1.22 *E*8	9.52 *E*7 (77.8%)	8.31 *E*7 (67.9%)
400	5.40 *E*8	4.18 *E*8 (77.4%)	3.70 *E*8 (68.6%)
800	2.41 *E*9	1.85 *E*9 (76.8%)	1.68 *E*9 (69.4%)

one gets

$$\begin{aligned} R(y) = \frac{y^T T y}{\|y\|_2^2} &= \frac{1}{1+\|w\|_2^2}(1, w^T)\begin{pmatrix} 1 & t^T \\ t & G \end{pmatrix}\begin{pmatrix} 1 \\ w \end{pmatrix} \\ &= \frac{1}{1+\|w\|_2^2}\big(1 + 2t^T w + w^T(G - \mu I + \mu I)w\big) \\ &= \frac{1 + t^T w + \mu\|w\|_2^2}{1+\|w\|_2^2}. \end{aligned}$$

This connection between Newton's method for the secular equation and the Rayleigh quotient was already pointed out by Trench [14].

Example 2. For each of the dimensions $n = 50$, 100, 200, 400, and 800 we treated 100 examples of the type described in Example 1. The Newton iteration was terminated when the relative increment $|f(\lambda)/(\lambda \cdot f'(\lambda))|$ was less than 10^{-6}. Column 2 of Table 2 contains the number of flops that was needed by the method of Cybenko and Van Loan. Columns 3 and 4 contain the improvements gained with the enhanced upper bound (2.6) at the start of the method and the improved upper bound by the Rayleigh quotient in subsequent steps, respectively.

3. Improvement by rational Hermitian interpolation. Although Newton's method for the secular equation $f(\lambda) = 0$ converges for any initial guess $\mu_0 \in (\lambda_1, \omega_1)$ and the convergence is quadratic, the global convergence behavior usually is not satisfactory. This is because f is a rational function where the root λ_1 we are looking for and the pole ω_1 can be very close to each other (cf. Table 1, Example 2). If μ_0 is close to ω_1 then the first steps of Newton's method can be extremely slow.

The convergence can be improved considerably if an iteration method is based on a better model of the rational function f than its tangent in Newton's method. The derivation of f by the elimination of the unknowns $x_2, \ldots, x_n$ is nothing else but the exact condensation of the eigenvalue problem $Tx = \lambda x$, where $x_2, \ldots, x_n$ are chosen to be slaves and x_1 is the only master. Using spectral information of the slave problem $(G - \mu I)v = 0$, the function f can be written as (cf. [8])

$$f(\lambda) = a_0 + a_1\lambda + \lambda^2 \sum_{j=1}^{n-1} \frac{\alpha_j^2}{\omega_j - \lambda},$$

where

$$a_0 := f(0) = t^T G^{-1} t - 1, \quad a_1 := f'(0) = 1 + \|G^{-1}t\|_2^2,$$

and α_j, $j = 1, \ldots, n-1$, are real numbers depending on the eigenvectors of G.

If we are given an approximation $\mu \in (0, \omega_1)$ we base a root finding method on the following rational substitute of f:

$$g(\lambda;\mu) := a_0 + a_1\lambda + \lambda^2 \frac{b}{c-\lambda}$$

with b and c determined by the Hermitian interpolation conditions

$$g(\mu;\mu) = f(\mu) \quad \text{and} \quad g'(\mu;\mu) = f'(\mu).$$

Rational approximations of the secular equation were already used by Dongarra and Sorensen [6] to compute eigenvalues in intermediate intervals (ω_j, ω_{j+1}) in a divide-and-conquer method.

Theorem 3.1 contains the basic properties of $g(\cdot;\mu)$.

THEOREM 3.1. *Let $\mu \in (0, \omega_1)$ and let*

$$g(\lambda;\mu) := a_0 + a_1\lambda + \lambda^2 \frac{b}{c-\lambda},$$

where

$$a_0 := t^T G^{-1} t - 1, \ a_1 := 1 + \|G^{-1}t\|_2^2,$$

and b and c are determined such that the interpolation conditions

$$g(\mu;\mu) = f(\mu), \quad g'(\mu;\mu) = f'(\mu)$$

are satisfied.

Then it holds that

(i)

$$b > 0 \quad \textit{and} \quad c > \mu,$$

whence $g(\cdot;\mu)$ is strictly monotonely increasing and strictly convex in $(0, c)$.

(ii)

$$g(\lambda_1;\mu) < 0 \quad \textit{for } \mu \neq \lambda_1.$$

Proof. (i): Let

$$f(\lambda) = a_0 + a_1\lambda + \lambda^2\phi(\lambda), \quad \phi(\lambda) := \sum_{j=1}^{n-1} \frac{\alpha_j^2}{\omega_j - \lambda}.$$

From the interpolation conditions one gets

$$\phi(\mu) = \frac{b}{c-\mu}, \quad \phi'(\mu) = \frac{b}{(c-\mu)^2}. \tag{3.1}$$

Since $\mu < \omega_j$ for every j

$$\phi(\mu) > 0, \quad \phi'(\mu) = \sum_{j=1}^{n-1} \frac{\alpha_j^2}{(\omega_j - \mu)^2} > 0,$$

and thus

$$\frac{b}{c-\mu} > 0, \quad \frac{b}{(c-\mu)^2} > 0,$$

yielding

$$b > 0 \quad \text{and} \quad c > \mu.$$

Differentiating g, one obtains for every $\lambda \in (0, \mu)$

$$g'(\lambda;\mu) = a_1 + 2\lambda\frac{b}{c-\lambda} + \lambda^2\frac{b}{(c-\lambda)^2} > 0,$$
$$g''(\lambda;\mu) = \frac{2bc^2}{(c-\lambda)^3} > 0.$$

Hence g is strictly monotonely increasing and strictly convex.

(ii): From

$$f(\lambda_1) = a_0 + a_1\lambda_1 + \lambda_1^2\phi(\lambda_1) = 0$$

we obtain

$$g(\lambda_1;\mu) = a_0 + a_1\lambda_1 + \lambda_1^2\frac{b}{c-\lambda_1} = \lambda_1^2\Big(\frac{b}{c-\lambda_1} - \phi(\lambda_1)\Big).$$

From (3.1) it follows that

$$b = \frac{\phi^2(\mu)}{\phi'(\mu)}, \quad c - \mu = \frac{\phi(\mu)}{\phi'(\mu)}.$$

Therefore,

$$\frac{b}{c-\lambda_1} = \frac{b}{c-\mu+(\mu-\lambda_1)} = \frac{\phi^2(\mu)}{\phi(\mu)+(\mu-\lambda_1)\phi'(\mu)}.$$

Since for $\mu \in (0, \omega_1)$

$$\begin{aligned}\phi(\mu) + \phi'(\mu)(\mu-\lambda_1) &= \sum_{j=1}^{n-1}\frac{\alpha_j^2}{\omega_j-\mu} + \sum_{j=1}^{n-1}\frac{\alpha_j^2(\mu-\lambda_1)}{(\omega_j-\mu)^2} \\ &= \sum_{j=1}^{n-1}\frac{\alpha_j^2(\omega_j-\lambda_1)}{(\omega_j-\mu)^2} > 0,\end{aligned}$$

the inequality $g(\lambda_1;\mu) < 0$ is equivalent to

$$\phi^2(\mu) < \phi(\lambda_1)\big(\phi(\mu) + \phi'(\mu)(\mu-\lambda_1)\big).$$

Hence, from

$$\begin{aligned}&\phi(\lambda_1)\big(\phi(\mu) + \phi'(\mu)(\mu-\lambda_1)\big) \\ &= \sum_{k=1}^{n-1}\frac{\alpha_k^2}{\omega_k-\lambda_1}\cdot\Big(\sum_{j=1}^{n-1}\frac{\alpha_j^2}{\omega_j-\mu} + (\mu-\lambda_1)\sum_{j=1}^{n-1}\frac{\alpha_j^2}{(\omega_j-\mu)^2}\Big) \\ &= \sum_{k=1}^{n-1}\frac{\alpha_k^2}{\omega_k-\lambda_1}\cdot\sum_{j=1}^{n-1}\frac{\alpha_j^2(\omega_j-\lambda_1)}{(\omega_j-\mu)^2} = \sum_{j,k=1}^{n-1}\frac{\alpha_j^2\alpha_k^2(\omega_j-\lambda_1)}{(\omega_k-\lambda_1)(\omega_j-\mu)^2}\end{aligned}$$

we obtain the following sequence of equivalent inequalities:

$$
\begin{aligned}
g(\lambda_1;\mu) < 0 &\Leftrightarrow \sum_{j,k=1}^{n-1} \Big(\frac{\alpha_j^2\alpha_k^2(\omega_j-\lambda_1)}{(\omega_k-\lambda_1)(\omega_j-\mu)^2} - \frac{\alpha_j^2\alpha_k^2}{(\omega_j-\mu)(\omega_k-\mu)} \Big) > 0 \\
&\Leftrightarrow \sum_{j,k=1}^{n-1} \frac{\alpha_j^2\alpha_k^2}{\omega_j-\mu} \cdot \Big(\frac{\omega_j-\lambda_1}{(\omega_k-\lambda_1)(\omega_j-\mu)} - \frac{1}{\omega_k-\mu} \Big) > 0 \\
&\Leftrightarrow \sum_{j,k=1}^{n-1} \frac{\alpha_j^2\alpha_k^2}{\omega_j-\mu} \cdot \frac{(\omega_j-\lambda_1)(\omega_k-\mu)-(\omega_k-\lambda_1)(\omega_j-\mu)}{(\omega_k-\lambda_1)(\omega_j-\mu)(\omega_k-\mu)} > 0 \\
&\Leftrightarrow \sum_{j,k=1}^{n-1} \frac{\alpha_j^2\alpha_k^2(\omega_k-\omega_j)(\mu-\lambda_1)}{(\omega_j-\mu)^2(\omega_k-\mu)(\omega_k-\lambda_1)} > 0 \\
&\Leftrightarrow \sum_{1\le j<k<n} \frac{\alpha_j^2\alpha_k^2(\mu-\lambda_1)}{(\omega_j-\mu)(\omega_k-\mu)} \cdot \Big(\frac{\omega_k-\omega_j}{(\omega_k-\lambda_1)(\omega_j-\mu)} + \frac{\omega_j-\omega_k}{(\omega_j-\lambda_1)(\omega_k-\mu)} \Big) > 0 \\
&\Leftrightarrow \sum_{1\le j<k<n} \frac{\alpha_j^2\alpha_k^2(\mu-\lambda_1)^2(\omega_k-\omega_j)^2}{(\omega_j-\mu)^2(\omega_k-\mu)^2(\omega_k-\lambda_1)(\omega_j-\lambda_1)} > 0.
\end{aligned}
$$

The last inequality is true since $\mu < \omega_j$ for $j = 1, \dots, n-1$ and $\mu \neq \lambda_1$. □

From Theorem 3.1 we deduce the following improvements of the method of Cybenko and Van Loan.

1. Let $\mu_n \in (\lambda_1, \omega_1)$ be a given approximation to λ_1. Then the function $g(\cdot;\mu_n)$ is strictly convex in the interval $(0, \mu_n)$. Since

$$g(\lambda_1;\mu_n) < 0 = f(\lambda_1) < f(\mu_n) = g(\mu_n;\mu_n),$$

$g(\mu;\mu_n)$ has exactly one solution: $\mu_{n+1} \in (\lambda_1, \mu_n)$.

From the convexity of $g(\cdot;\mu_n)$ we obtain

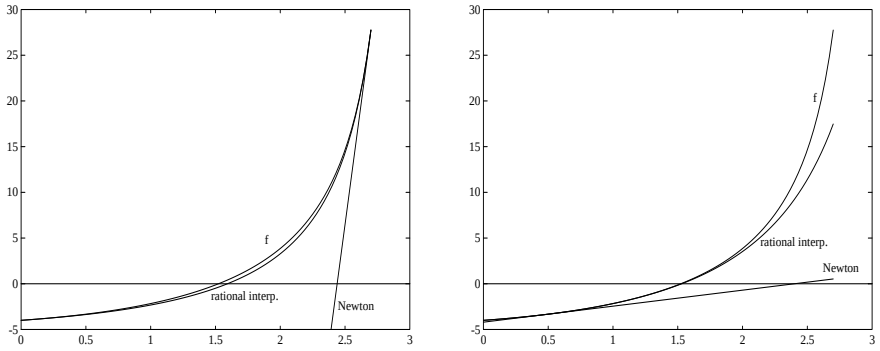
$$
\begin{aligned}
g(\mu;\mu_n) &> g(\mu_n;\mu_n) + g'(\mu_n;\mu_n)(\mu-\mu_n) \\
&= f(\mu_n) + f'(\mu_n)(\mu-\mu_n) \quad \text{for every } \mu \in (\lambda_1, \mu_n),
\end{aligned}
$$

and thus μ_{n+1} is always a better approximation to λ_1 than the Newton iterate with initial guess μ_n. Hence, for $\mu_0 \in (\lambda_1, \omega_1)$ the method which defines μ_{n+1} as the unique root of the rational Hermitian interpolation $g(\mu;\mu_n)$ in $(0, \mu_n)$ converges monotonely decreasing to λ_1, and it is guaranteed to be faster than Newton's method (cf. Fig. 1(a) for the function $f(\lambda) = -4 + \lambda + \lambda^2 \sum_{j=1}^{5} 1/(j*(2+j-\lambda))$ and $\mu_n = 2.7$).

Notice that the costs of Newton's method and the method defined above are nearly identical. One must solve one Yule–Walker system ($2n^2$ flops) and evaluate two inner products to obtain $f(\mu_n)$ and $f'(\mu_n)$. The determination of b and c and the solution of a quadratic equation to obtain μ_{n+1} need only $O(1)$ flops and can be neglected.

2. If in the bisection process a test parameter μ_n is contained in the interval $(0, \lambda_1)$ then it follows from $g(\lambda_1;\mu_n) < 0$ that the unique root $\tilde{\mu} \in (\mu_n, c)$ of $g(\cdot;\mu_n)$ is an upper bound of λ_1. Since $g(\cdot;\mu_n)$ is strictly convex we obtain in the same way as in the case $\mu \in (\lambda_1, \omega_1)$ that

$$\lambda_1 < \tilde{\mu} < \mu_n - \frac{f(\mu_n)}{f'(\mu_n)} =: \hat{\mu}.$$



(a) (b)

FIG. 1.

Hence, we replace the current upper bound of λ_1 by $\tilde{\mu}$ whenever this improves the inclusion of λ_1 (cf. Fig. 1(b) for f as above and $\mu_n = 0.5$).

If the test parameter μ_n is close to λ_1 then by the quadratic convergence $\tilde{\mu}$ will be an excellent approximation to λ_1. Therefore, it is reasonable to choose $\tilde{\mu}$ as the next test parameter if $\tilde{\mu}$ replaces the current upper bound. However, if the gap between λ_1 and ω_1 is very small, then it may happen that the method bounces between good upper bounds μ_{n+1} of λ_1 which are produced as roots of rational interpolates $g(\cdot;\mu_n)$ and are contained in (ω_1,∞) and lower bounds $\omega_{n+2} := 0.5(\omega_n + \omega_{n+1})$ which are obtained in the bisection process. For instance, in the second problem of Example 1 one obtains the following sequence of test parameters:

$$\begin{array}{rlrl}
\mu_6 = 2.1901E-6 & < \lambda_1 < & & < \omega_1 < \quad \mu_7 = 3.4895E-6 \\
\mu_8 = 2.8398E-6 & < \lambda_1 < & & < \omega_1 < \quad \mu_9 = 3.4892E-6 \\
\mu_{10} = 3.1645E-6 & < \lambda_1 < & & < \omega_1 < \quad \mu_{11} = 3.4890E-6 \\
\mu_{12} = 3.3268E-6 & < \lambda_1 < & & < \omega_1 < \quad \mu_{13} = 3.4888E-6 \\
\mu_{14} = 3.4078E-6 & < \lambda_1 < & & < \omega_1 < \quad \mu_{15} = 3.4885E-6 \\
\mu_{16} = 3.4481E-6 & < \lambda_1 < & & < \omega_1 < \quad \mu_{17} = 3.4875E-6 \\
\mu_{18} = 3.4678E-6 & < \lambda_1 < & & < \omega_1 < \quad \mu_{19} = 3.4843E-6 \\
\mu_{20} = 3.4761E-6 & < \lambda_1 < & \mu_{21} = 3.4780E-6 & < \omega_1
\end{array}$$

To break a tie like this we introduced the following modification: determining the coefficients b and c of the rational function, we have already evaluated $f(\mu_n)$ and $f'(\mu_n)$. Hence, along with the root $\tilde{\mu}$ of $g(\cdot;\mu_n)$ we can obtain the Newton iterate $\hat{\mu}$ at a negligible cost. Since for a sufficiently good approximation μ_n to λ_1 both $\tilde{\mu}$ and $\hat{\mu}$ are approximations to λ_1 with error

$$\tilde{\mu} - \lambda_1 = O(|\lambda_1 - \mu_n|^2), \quad \hat{\mu} - \lambda_1 = O(|\lambda_1 - \mu_n|^2)$$

the relative difference $(\hat{\mu} - \tilde{\mu})/\tilde{\mu}$ is an indicator whether μ_n is close to λ_1 or not. For

$$\frac{\hat{\mu} - \tilde{\mu}}{\tilde{\mu}} < 0.01$$

we continued the bisection process with the test parameter $\mu_{n+1} = \tilde{\mu}$; otherwise we chose $\mu_{n+1} := 0.1\mu + 0.9\tilde{\mu}$.

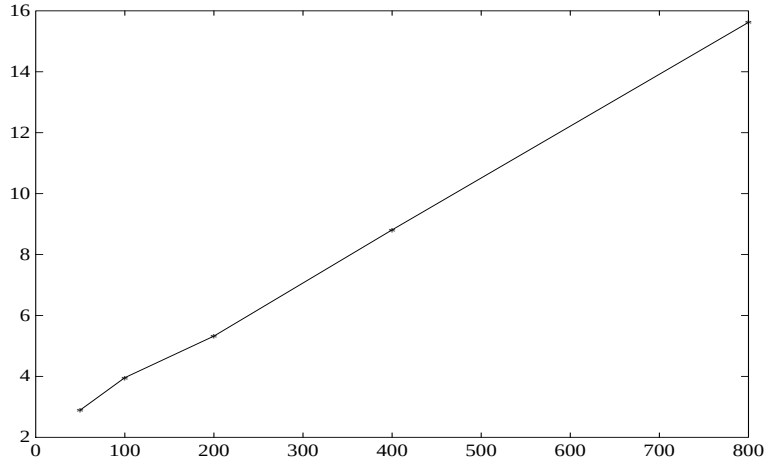


FIG. 2.

In problem 2 of Example 1 we obtained with this modification

$$\begin{array}{llll} \mu_6 = 2.190116E-6 & < \lambda_1 < & & < \omega_1 \\ \mu_7 = 3.359595E-6 & < \lambda_1 < & & < \omega_1 \\ \mu_8 = 3.475845E-6 & < \lambda_1 < & \mu_9 = 3.478148E-6 & < \omega_1 \end{array}$$

3. If $\mu_n \in (0, \lambda_1)$ for some $n \in \mathbb{N}$ then the matrix $T_n := T - \mu_n I$ is a symmetric and positive-definite Toeplitz matrix. Hence, the method described above applies to T_n. The secular equation of T_n is

$$f_n(\lambda) := f(\mu_n) + f'(\mu_n)(\lambda - \mu_n) + (\lambda - \mu_n)^2 \sum_{j=1}^{n-1} \frac{\tilde{\alpha}_j^2}{\omega_j - \lambda}.$$

It is easily seen that the root of the rational interpolation of f_n for any test parameter $\mu \in (\mu_n, \omega_1)$ is a better approximation to λ_1 than the root of $g(\cdot; \mu)$.

4. The considerations above indicate that the benefit of a small test parameter μ is much bigger than that of a large parameter μ. All that is gained from Durbin's algorithm for a parameter $\mu > \omega_1$ is the fact that $\mu > \omega_1$. For $\mu \in (0, \lambda_1)$, however, we obtain a potential new upper bound $\tilde{\mu}$ as well as a shift in the sense of item 3. For $\mu \in (\lambda_1, \omega_1)$ one can even start the quadratically convergent phase of the algorithm.

Hence, it is not reasonable to start the algorithm with the initial upper bound $R(y)$, $y = T^{-1}e^1$ as the first test parameter, but with a suitable reduction of this value instead.

Figure 2 contains the graph of the mean value of the quotients of $R(y)$ and λ_1 for the 100 problems considered in Example 2 for the dimensions $n = 50, 100, 200, 400, 800$. It indicates that these quotients grow linearly with the dimension of the problem. A least squares fit yielded

$$R(y) \approx (2.0911 + 0.0169n) \cdot \lambda_1.$$

Cybenko and Van Loan terminated Newton's iteration for f if the relative increment $(\mu_{n+1} - \mu_n)/\mu_n$ was less than a prescribed tolerance δ (and so did we in

Example 2). By the quadratic convergence of Newton's method this term is usually a reasonable estimate of the relative error of μ_{n+1}.

Since $f'''(\lambda) \geq 0$ for every $\lambda \in (0, \omega_1)$ it offers no problems, however, to obtain an exact and realistic error bound of μ_{n+1} by Hermitian interpolation.

LEMMA 3.2. *Let $\mu_n \in (\lambda_1, \omega_1)$ be given and let $\hat{\mu} \in [0, \lambda_1)$ be a lower bound of λ_1 (which was obtained in a previous step). Let p be the unique quadratic polynomial satisfying the interpolation conditions*

$$p(\hat{\mu}) = f(\hat{\mu}), \quad p'(\hat{\mu}) = f'(\hat{\mu}), \quad p(\mu_n) = f(\mu_n).$$

Then p has a unique root κ in the interval $[\hat{\mu}, \mu_n]$ and

$$\kappa \leq \lambda_1.$$

Proof. It is easily seen that Green's function g of the boundary value problem

$$L[u] := u''', \ u(\hat{\mu}) = 0, \ u'(\hat{\mu}) = 0, \ u(\mu_n) = 0$$

is negative on $(\hat{\mu}, \mu_n) \times (\hat{\mu}, \mu_n)$.

Let $v(x) := p(x) - f(x)$. Then v satisfies the boundary conditions $v(\hat{\mu}) = 0$, $v'(\hat{\mu}) = 0$, and $v(\mu_n) = 0$. Hence, from $f'''(x) > 0$ on $[\hat{\mu}, \mu_n]$ we get

$$v(x) = \int_{\hat{\mu}}^{\mu_n} g(x,t) v'''(t)\, dt = -\int_{\hat{\mu}}^{\mu_n} g(x,t) f'''(t)\, dt > 0, \ x \in (\hat{\mu}, \mu_n),$$

i.e., $p(x) > f(x)$ on $(\hat{\mu}, \mu_n)$, and therefore the unique root κ of p in $[\hat{\mu}, \mu_n]$ is a lower bound of the unique root λ_1 of f. ∎

4. A MATLAB program. In the following we give a MATLAB program for the determination of the smallest eigenvalue of a symmetric and positive-definite Toeplitz matrix based on the considerations above.

Let $[w, where] = durbin(\mu)$ denote a function which for a given test parameter μ returns the integer variable

$$\text{where} = \begin{cases} 0 & \text{if } \mu \in (0, \lambda_1), \\ 1 & \text{if } \mu \in [\lambda_1, \omega_1), \\ 2 & \text{if } \mu \in (\omega_1, \infty), \end{cases}$$

and for $\mu \in (0, \omega_1)$ additionally the solution w of the Yule–Walker system $(G - \mu I)w = -t$. Notice that in the case $\mu > \omega_1$ the Durbin algorithm is terminated as soon as a negative diagonal element E_j is detected. Hence, for $\mu \in (0, \omega_1)$ a call of Durbin needs $2n^2$ flops; for $\mu > \omega_1$ it needs less than $2n^2$ flops.

Let $\tilde{\mu} = fracroot(\mu; \alpha, a_0, a_1)$ return the unique root in (α, μ) of the rational function

$$g(\lambda) = a_0 + a_1(\lambda - \alpha) + (\lambda - \alpha)^2 \frac{b}{c - \lambda},$$

which satisfies the Hermitian interpolation conditions

$$g(\alpha) = f(\alpha) =: a_0, \ g'(\alpha) = f'(\alpha) =: a_1, \ g(\mu) = f(\mu), \ g'(\mu) = f'(\mu),$$

and let $\kappa = quadroot(\mu; \alpha, a_0, a_1)$ return the unique root in (α, μ) of the quadratic polynomial p satisfying the Hermitian interpolation conditions

$$p(\alpha) = f(\alpha) =: a_0,\ p'(\alpha) = f'(\alpha) =: a_1,\ p(\mu) = f(\mu).$$

Then the final algorithm reads as follows.

```
[w,where]=durbin(0);
a0=-(1+w'*t);
a1=1+w'*w;
beta=-a0/a1;
alpha=0;
mu=beta/(2.0911+0.0169*n)
h=1;
while abs(h) > mu*(1.e-6)
  [w,where]=durbin(mu);
  if where == 2
    beta=mu;
    mu=0.5*(alpha+beta);
    h=beta-mu;
  elseif where == 0
    f=mu-1-w'*t;
    f_prime=1+w'*w;
    lambda_1=mu-f/f_prime;
    lambda_2=fracroot(mu;alpha;a0,a1);
    alpha=mu;
    a0=f;
    a1=f_prime;
    if lambda_2 < beta
      beta=lambda_2;
      if (lambda_1-lambda_2)/lambda_2 < 0.01
        mu=lambda_2;
      else
        mu=0.1*alpha+0.9*beta;
        end;
    else
      mu=0.5*(alpha+beta);
      end;
    h=beta-alpha;
  else
    f=mu-1-w'*t;
    f_prime=1+w'*w;
    kappa=quadroot(mu;alpha;a0,a1);
    mu=fracroot(mu;alpha;a0,a1);
    h=mu-kappa;
  end;
```

5. Numerical experiments. With the algorithm of the preceding section we solved the test problems of Example 2. We terminated the iteration if according to the error bound of Lemma 3.2 the relative error was less than 10^{-6}. Column 2 of Table 3 contains the numbers of flops needed with the method of Cybenko and Van Loan (which are a little smaller than the numbers in Table 2 since we used the error

TABLE 3
Number of flops for 100 *test examples.*

dimension	Cybenko–Van Loan	rational approximation	initial guess from least squares fit	modified initial guess
50	6.33 *E*6	3.21 *E*6 (50.7%)	2.58 *E*6 (40.8%)	2.62 *E*6 (41.4%)
100	2.68 *E*7	1.35 *E*7 (50.2%)	1.04 *E*7 (38.8%)	1.06 *E*6 (39.4%)
200	1.20 *E*8	5.77 *E*7 (48.2%)	4.26 *E*7 (35.6%)	4.31 *E*7 (36.0%)
400	5.34 *E*8	2.58 *E*8 (48.4%)	1.82 *E*8 (34.0%)	1.84 *E*8 (34.6%)
800	2.37 *E*9	1.22 *E*9 (51.5%)	8.42 *E*8 (35.6%)	8.34 *E*8 (35.2%)

TABLE 4
Average number of linear systems for 100 *test examples.*

dimension	rational app.	Lanczos method
50	4.77	6.59
100	5.01	6.91
200	5.13	7.37
400	5.64	7.61
800	6.57	7.44

bound from Lemma 3.2). Columns 3, 4, and 5 contain the numbers of flops for the method given above where we chose as initial test parameter the initial upper bound $R(y)$ where $y := T^{-1}e^1$, the approximation $\tilde{\mu}_0 := R(y)/(2.0911 + 0.0169n)$ to λ_1 suggested by the least squares fit, and the lower bound $\hat{\mu}_0 := R(y)/(4+0.02n)$ of that value. It turns out that the rough underestimation $\hat{\mu}_0$ of the approximation $\tilde{\mu}_0$ does not influence the performance of the algorithm very much.

We compared our method to Lanczos's method applied to T^{-1}. Since for this method the error bound of Lemma 3.2 does not apply we terminated the iteration if the eigenvalue approximation was less than $\mu * (1 + 1.E - 6)$, where μ denotes the approximate value with relative error less than 10^{-6} obtained with our method.

For the same test problems as before Table 4 shows the average number of linear systems that must be solved in each step in the new method (with initial guess $R(y)/(2.0911 + 0.0169n)$) and the Lanczos method.

Notice that the accuracy requirements for the Lanczos method are weaker than those for the rational Hermitian interpolation. Moreover, we only have to solve Yule–Walker systems, whereas in the Lanczos method in every step one must solve a general linear system $Tu = v$. This requires $2n^2$ flops in each step to solve two triangular systems and additionally $2n^2$ flops to compute the decomposition corresponding to (2.5) and n^2 storage locations.

6. Concluding remarks. We have presented an algorithm for the computation of the minimum eigenvalue of a symmetric and positive-definite Toeplitz matrix which improves the method of Cybenko and Van Loan considerably and which is superior to the Lanczos method. Realistic error bounds are obtained at a negligible cost. In our numerical tests we used Durbin's algorithm to solve Yule–Walker systems and to determine the diagonal matrix in the decomposition (2.5). This information can be gained from superfast Toeplitz solvers (cf. [1], [2], [4]) as well. Hence, the computational complexity can be reduced to $O(n \log^2 n)$ operations.

REFERENCES

[1] G. S. Ammar and W. B. Gragg, *The generalized Schur algorithm for the superfast solution of Toeplitz systems*, in Rational Approximation and its Applications in Mathematics and Physics, J. Gilewicz, M. Pindor, and W. Siemaszko, eds., Lecture Notes in Mathematics 1237, Berlin, 1987, pp. 315–330.

[2] G. S. Ammar and W. B. Gragg, *Numerical experience with a superfast real Toeplitz solver*, Linear Algebra Appl., 121 (1989), pp. 185–206.

[3] G. Cybenko and C. Van Loan, *Computing the minimum eigenvalue of a symmetric positive definite Toeplitz matrix*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 123–131.

[4] F. de Hoog, *A new algorithm for solving Toeplitz systems of equations*, Linear Algebra Appl., 88/89 (1987), pp. 123–138.

[5] A. Dembo, *Bounds on the extreme eigenvalues of positive definite Toeplitz matrices*, IEEE Trans. Inform. Theory, 34 (1988), pp. 352–355.

[6] J. J. Dongarra and D. C. Sorensen, *A fully parallel algorithm for the symmetric eigenvalue problem*, SIAM J. Sci. Stat. Comput., 8 (1987), pp. s139–s154.

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

[8] T. Hitziger, W. Mackens, and H. Voss, *A condensation-projection method for generalized eigenvalue problems,* in High Performance Computing in Engineering 1, H. Power and C.A. Brebbia, eds., Computational Mechanics Publications, Southampton, 1995, pp. 239–282.

[9] Y. H. Hu and S.-Y. Kung, *Toeplitz eigensystem solver*, IEEE Trans. Acoust. Speech Signal Process., 33 (1985), pp. 1264–1271.

[10] T. Huckle, *Computing the minimum eigenvalue of a symmetric positive definite Toeplitz matrix with spectral transformation Lanczos method*, in Numerical Treatment of Eigenvalue Problems, Vol. 5, J. Albrecht, L. Collatz, P. Hagedorn, and W. Velte, eds., Birkhäuser-Verlag, Basel, Switzerland, 1991, pp. 109–115.

[11] T. Huckle, *Circulant and skewcirculant matrices for solving Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 767–777.

[12] F. Noor and S. D. Morgera, *Recursive and iterative algorithms for computing eigenvalues of Hermitian Toeplitz matrices*, IEEE Trans. Signal Process., 41 (1993), pp. 1272–1280.

[13] V. F. Pisarenko, *The retrieval of harmonics from a covariance function*, Geophys. J. R. Astr. Soc., 33 (1973), pp. 347–366.

[14] W. F. Trench, *Numerical solution of the eigenvalue problem for Hermitian Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 135–146.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 3, pp. 535–551, July 1997

002

ESTIMATING THE ATTAINABLE ACCURACY OF RECURSIVELY COMPUTED RESIDUAL METHODS*

ANNE GREENBAUM†

Abstract. Many conjugate gradient-like methods for solving linear systems $Ax = b$ use recursion formulas for updating residual vectors instead of computing the residuals directly. For such methods it is shown that the difference between the actual residuals and the updated approximate residual vectors generated in finite precision arithmetic depends on the machine precision ϵ and on the maximum norm of an iterate divided by the norm of the true solution. It is often observed numerically, and can sometimes be proved, that the norms of the updated approximate residual vectors converge to zero or, at least, become orders of magnitude smaller than the machine precision. In such cases, the actual residual norm reaches the level $\epsilon\|A\|\|x\|$ times the maximum ratio of the norm of an iterate to that of the true solution. Using exact arithmetic theory to bound the size of the iterates, we give a priori estimates of the size of the final residual for a number of algorithms.

Key words. iterative methods, accuracy, finite precision arithmetic

AMS subject classifications. 65F10, 65F15

PII. S0895479895284944

1. Introduction.

Many iterative methods for solving a linear system $Ax = b$ start with an initial guess x^0 for the solution, compute the initial residual $r^0 = b - Ax^0$, and then generate updated approximations x^k and residuals r^k, $k = 1, 2, \ldots$, according to the following formulas:

$$x^k = x^{k-1} + a_{k-1}p^{k-1}, \qquad r^k = r^{k-1} - a_{k-1}Ap^{k-1}. \tag{1}$$

Here p^{k-1} is some direction vector and a_{k-1} some coefficient. The vector r^k is updated instead of computed directly as $b - Ax^k$. We are particularly interested in methods in which the coefficients and direction vectors are computed in terms of the updated vectors r^k.

Often a preconditioner M, which approximates A but is easier to invert than A, is used to accelerate convergence. In this case a vector z^k is computed by solving $Mz^k = r^k$. The solution z^k may be used in determining the new coefficient and direction vector, but it does not alter the general formulas in (1). It is these formulas that we will analyze in finite precision arithmetic.

Examples of algorithms that are usually implemented in the form (1) include the following.

- The steepest descent method for symmetric positive definite problems. Here the vector p^{k-1} is equal to z^{k-1}, and the coefficient a_{k-1} is chosen to minimize the A^{-1}-norm of r^k, $\|r^k\|_{A^{-1}} \equiv \langle r^k, A^{-1}r^k\rangle^{1/2}$, in the direction p^{k-1}.
- The conjugate gradient (CG) method for symmetric positive definite problems [16, 5]. Here the coefficient a_{k-1} is chosen to minimize the A^{-1}-norm of r^k in the direction p^{k-1}, and the direction vectors form an A-orthogonal basis for the Krylov space $[z^0, M^{-1}Az^0, \ldots, (M^{-1}A)^{k-1}z^0]$.
- BCG and CGS for general nonsymmetric problems [9, 18]. These methods use two sets of recurrences, one involving A, like formulas (1), and another involving

* Received by the editors April 21, 1995; accepted for publication (in revised form) by D. P. O'Leary July 2, 1996. This work was supported in part by DOE contract DEFG0288ER25053.
http://www.siam.org/journals/simax/18-3/28494.html

† Courant Institute of Mathematical Sciences, 251 Mercer St., New York, NY 10012 (greenbau@nyu.edu).

A^T. They may break down, even in exact arithmetic, due to the breakdown of the underlying two-sided Lanczos recurrence or the nonexistence of the BCG iterate at certain steps. In either case look-ahead steps may be used to avoid the steps at which the Lanczos vectors or BCG iterates are undefined [11]. We will consider only problems in which look-ahead steps are not required and x^k and r^k are updated as in (1).

• CGNR and CGNE for nonsymmetric problems [16, 6]. These methods are like the CG method for the normal equations $A^TAx = A^Tb$ or $AA^Ty = b,\ x = A^Ty$. In CGNR, the coefficient a_{k-1} is chosen to minimize the 2-norm of r^k, while in CGNE it minimizes the $(A^TA)^{-1}$-norm of r^k, $\langle A^{-1}r^k, A^{-1}r^k\rangle^{1/2}$. While the Krylov space over which the minimization is performed is the same as that used by CG applied to the normal equations, one does not actually form the normal equations, and approximate solutions and residual vectors are still computed by formulas (1).

A number of other methods are *not* usually implemented in the form (1), but, with some modifications, they could be. These include the following.

• Stationary iterative methods, such as Jacobi, Gauss–Seidel, SOR, etc. For these methods, residuals are usually computed directly. They could be updated as in (1), but there appears to be no particular advantage in doing so.

• ORTHOMIN and ORTHODIR for nonsymmetric problems [20, 23]. In standard implementations, x^k is computed as in (1), but r^k is set to $r^{k-1} - a_{k-1}q^{k-1}$, where q^{k-1} is a vector that is equal to Ap^{k-1} in exact arithmetic but might differ from this in finite precision arithmetic. The vector q^{k-1} could be explicitly set to Ap^{k-1}, but this would require an extra matrix–vector multiplication at each iteration.

• QMR for nonsymmetric problems [11]. Like BCG, this method uses two sets of recurrences, one involving A and another involving A^T, and like BCG, it can break down if the underlying two-sided Lanczos process breaks down. Such breakdowns can be avoided by using look-ahead steps. A number of different QMR implementations have been proposed [11, 12]. The one given in [2] updates x^k as $x^k = x^{k-1} + d^{k-1}$, but sets $r^k = r^{k-1} - s^{k-1}$, where s^{k-1} is a vector that would be equal to Ad^{k-1} in exact arithmetic but might differ from this in finite precision arithmetic.

• Bi-CGSTAB for nonsymmetric problems [19]. This is another modification of BCG designed to smooth the erratic convergence behavior of BCG. The implementation given in [2] updates x^k from x^{k-1} and two different vectors, while setting r^k to r^{k-1} minus A times the appropriate linear combination of these two vectors. The analysis of recurrences of this form should be very similar to that of (1).

In section 2 we consider the implementation of formulas (1) in finite precision arithmetic. A bound is given on the difference between the true residuals $b - Ax^k$ and the updated vectors r^k. Specifically, it is shown that

$$\frac{\|b - Ax^k - r^k\|}{\|A\|\ \|x\|} \leq \epsilon\ O(k)\ (1 + \max_{j\leq k} \|x^j\|/\|x\|), \tag{2}$$

where ϵ is the machine precision. It is argued that the last term on the right-hand side of (2)—the maximum norm of an iterate divided by the norm of the true solution—plays an important role in determining the size of the quantity on the left-hand side of (2).

It is often observed numerically, and in some cases can be proved, that the updated vectors r^k converge to zero as $k \to \infty$ or, at least, that their norms become many orders of magnitude smaller than the machine precision. For certain algorithms, such as the steepest descent method and some implementations of the CG method, this can

be proved under reasonable assumptions about the condition number of the matrix just by considering the effect of individual steps [3, 21, 22]. Roughly, this is because the coefficients are chosen to minimize some norm of r^k at each step. We will not attempt to prove this here but will demonstrate numerically that the updated approximate residual vectors often become much smaller than the machine precision. In such cases, the right-hand side of (2) gives a reasonable estimate of the best attainable actual residual. Note that this analysis does *not* deal with the rate of convergence of iterative methods in finite precision arithmetic but only with the level of accuracy attainable if the iteration is carried out for sufficiently many steps and assuming that the vectors r^k become tiny.

The size of the maximum iterate divided by the norm of the true solution in (2) can be estimated for various algorithms, assuming exact arithmetic, and these estimates often hold in finite precision arithmetic as well. Using this combination of rigorous finite precision analysis of formulas (1) and exact arithmetic estimates of the size of the iterates, we consider specific algorithms and give numerical examples in section 3. For methods in which the 2-norm of the error decreases monotonically, such as the (unpreconditioned) steepest descent method and the (unpreconditioned) CG method for symmetric positive definite problems and the CGNE and (unpreconditioned) CGNR methods for nonsymmetric problems, the last term in (2) is bounded by $2 + \|x^0\|/\|x\|$. Hence, if the number of steps required to reduce the norm of r^k below $O(\epsilon)\|A\|\|x\|$ is not too large, then these algorithms will return an approximate solution for which the relative residual is of order ϵ. The relative error, $\|x - x^k\|/\|x\|$, is bounded by $\kappa(A)\epsilon$, where $\kappa(\cdot)$ is the condition number. For methods in which the 2-norm of the error may grow, but some other norm, say, the B-norm of the error, decreases monotonically, we can show only that the last factor in (2) is bounded by $1 + \kappa^{1/2}(B)(1 + \|x^0\|/\|x\|)$, suggesting a relative residual of order $\kappa^{1/2}(B)\epsilon$ and a relative error of order $\kappa(A)\kappa^{1/2}(B)\epsilon$. For BCG and CGS and other methods that do not necessarily reduce any standard error norm, the last factor in (2) cannot be bounded a priori. Such methods may occasionally fail to generate an accurate approximate solution for even a well-conditioned problem, although the updated vectors r^k may become tiny. An example of this is given.

Several of the previously listed algorithms have been analyzed by others. Higham and Knight [15] analyzed the effects of finite precision arithmetic on stationary iterative methods when the residuals are computed directly instead of updated. Wozniakowski [22] considered the steepest descent method and a special version of the CG algorithm, again with directly computed residuals, and gave bounds on the ultimately attainable accuracy in finite precision arithmetic. We will argue later that for the CG algorithm it is better (in terms of rate of convergence) to use the update formula in (1) than to compute residuals directly. The work most closely related to our own is that of Bollen [3], who also analyzed the effects of finite precision arithmetic on iterative methods having a form similar to (1). Bollen showed, under certain assumptions, that the vectors r^k converge at least linearly to a small value. Finally, results similar to (2) have been observed numerically by van der Vorst [19]. Van der Vorst noted that an increase in the 2-norm of the residual at intermediate steps leads to a corresponding increase in the size of the final residual. Inequality (2) shows that it is not really the size of intermediate residuals that is of importance but the size of the iterates. We give an example in which the residual remains small but intermediate iterates grow, causing a loss of accuracy in the final solution.

It should also be noted that the quantity $\|b - Ax^k\|/\|b\|$, which is often the value

actually monitored, may be much larger than $\|b-Ax^k\|/(\|A\|\,\|x\|)$ if $\|b\| << \|A\|\,\|x\|$. Our results deal only with the latter expression, which we refer to as the relative residual norm. While the relative residual norm may be small even though the size of an iterate is large, the reverse cannot occur:

$$\frac{\|b-Ax^k\|}{\|A\|\,\|x\|} \le \frac{\|b\|+\|A\|\,\|x^k\|}{\|A\|\,\|x\|} \le 1+\frac{\|x^k\|}{\|x\|}.$$

2. Finite precision implementation of formulas (1). We assume the following model of floating point arithmetic on a machine with unit roundoff ϵ:

$$\text{fl}(a \pm b) = a(1+\epsilon_1) \pm b(1+\epsilon_2), \quad |\epsilon_1|, |\epsilon_2| \le \epsilon, \tag{3}$$

$$\text{fl}(a \text{ op } b) = (a \text{ op } b)(1+\epsilon_3), \quad |\epsilon_3| \le \epsilon, \quad \text{op} = *, /. \tag{4}$$

This model is valid even for machines that do not use a guard digit in addition and subtraction.

Under this model, we have the following standard results for operations involving n-vectors v and w and a number a:

$$\|av - \text{fl}(av)\| \le \epsilon\,\|av\|, \tag{5}$$

$$\|v+w-\text{fl}(v+w)\| \le \epsilon\,(\|v\|+\|w\|), \tag{6}$$

$$|\langle v,w\rangle - \text{fl}(\langle v,w\rangle)| \le n\,(\epsilon+O(\epsilon^2))\,\|v\|\,\|w\|. \tag{7}$$

With these rules, we now consider the implementation of formulas (1) in finite precision arithmetic. Throughout this section, x^k, r^k, a_{k-1}, and p^{k-1} will always denote the quantities *actually computed*. This should cause no confusion since we will never have occasion to refer to the quantities that would be generated in exact arithmetic. To keep the exposition as simple as possible, we will express terms involving ϵ^2 or higher powers of ϵ as $O(\epsilon^2)$ when there are similar terms of order ϵ present. When formulas (1) are implemented in finite precision arithmetic, the computed iterates satisfy

$$x^k = x^{k-1} + a_{k-1}p^{k-1} + \xi^k, \tag{8}$$

$$r^k = r^{k-1} - a_{k-1}Ap^{k-1} + \eta^k, \tag{9}$$

where

$$\|\xi^k\| \le \epsilon\,\|x^{k-1}\| + (2\epsilon+\epsilon^2)\,\|a_{k-1}p^{k-1}\| \tag{10}$$

and

$$\|\eta^k\| \le \epsilon\,\|r^{k-1}\| + (2\epsilon+\epsilon^2)\,\|a_{k-1}Ap^{k-1}\| + (1+\epsilon)^2\,\|a_{k-1}d^{k-1}\|, \tag{11}$$

where $\mathrm{fl}(Ap^{k-1}) = Ap^{k-1} + d^{k-1}$. Inequalities (10)–(11) follow from a straightforward application of rules (5)–(6). The size of the term d^{k-1} depends on the accuracy of the matrix–vector multiplication routine, and we will assume that d^{k-1} satisfies

$$\|d^{k-1}\| \le c\ \epsilon\ \|A\|\ \|p^{k-1}\|. \tag{12}$$

It can be shown from estimate (7) that if A is an n by n matrix with at most m nonzeros in any row and if the matrix–vector product is computed in the standard way, then $c = mn^{1/2}$. We will not make this assumption here, however, because sometimes the matrix A is not stored explicitly and different procedures are used to compute the matrix–vector product.

Multiplying equation (8) by A and subtracting from b gives a recurrence for the true residual $b - Ax^k$. Subtracting from this the recurrence (9) for r^k, we find

$$\begin{aligned} b - Ax^k - r^k &= (b - Ax^{k-1} - r^{k-1}) - A\xi^k - \eta^k \\ &= (b - Ax^0 - r^0) - \sum_{j=1}^{k} (A\xi^j + \eta^j). \end{aligned}$$

Taking norms on both sides and dividing by $\|A\|\ \|x\|$ gives

$$\frac{\|b - Ax^k - r^k\|}{\|A\|\ \|x\|} \le \frac{\|b - Ax^0 - r^0\|}{\|A\|\ \|x\|} \ +\ \sum_{j=1}^{k} \left(\frac{\|\xi^j\|}{\|x\|} + \frac{\|\eta^j\|}{\|A\|\ \|x\|} \right). \tag{13}$$

Equality can hold in (13) for certain vectors ξ^j and η^j whose norms satisfy (10)–(11).

Since the initial vector r^0 is computed directly, the first term on the right-hand side of (13) is easily bounded using rule (6) and expression (12) for the accuracy of the matrix–vector multiplication routine:

$$\|b - Ax^0 - r^0\| \le \epsilon\ ((1 + c)\ \|A\|\ \|x^0\| + \|b\|) + c\ \epsilon^2\ \|A\|\ \|x^0\|,$$

and since $\|b\| \le \|A\|\ \|x\|$, we can write

$$\frac{\|b - Ax^0 - r^0\|}{\|A\|\ \|x\|} \le \epsilon(1 + c)\frac{\|x^0\|}{\|x\|} + \epsilon + c\epsilon^2\frac{\|x^0\|}{\|x\|}. \tag{14}$$

The following lemma bounds the other terms in (13).

LEMMA 2.1. *Define*

$$\Theta_k \equiv \max_{j \le k} \|x^j\| / \|x\|. \tag{15}$$

Assume that $1 - 2\epsilon - \epsilon^2 > 0$. *Then the terms on the right-hand side of* (13) *satisfy*

$$\sum_{j=1}^{k} \frac{\|\xi^j\|}{\|x\|} \le (5\ \epsilon + O(\epsilon^2))\ k\ \Theta_k, \tag{16}$$

$$\sum_{j=1}^{k} \frac{\|\eta^j\|}{\|A\|\ \|x\|} \le (\epsilon + O(\epsilon^2))\ k\ (1 + (5 + 2c)\ \Theta_k). \tag{17}$$

Proof. From (8), we can write

$$a_{j-1}p^{j-1} = x^j - x^{j-1} - \xi^j, \tag{18}$$

and substituting this expression into (10) gives

$$\|\xi^j\| \le \epsilon\|x^{j-1}\| + (2\epsilon + \epsilon^2)(\|x^j\| + \|x^{j-1}\| + \|\xi^j\|).$$

Using the assumption $1 - 2\epsilon - \epsilon^2 > 0$, this can be written in the form

$$\|\xi^j\| \le \epsilon\ (3\|x^{j-1}\| + 2\|x^j\|) + O(\epsilon^2)\ (\|x^{j-1}\| + \|x^j\|). \tag{19}$$

From this, (16) follows by bounding the sum on the left in (16) by k times the maximum term.

Using (12), the third term in expression (11) for η_j can be bounded by

$$(1+\epsilon)^2\ \|a_{j-1}d^{j-1}\| \le c\ \epsilon\ (1+\epsilon)^2\ \|A\|\ \|a_{j-1}p^{j-1}\|,$$

and expressing $a_{j-1}p^{j-1}$ as in (18) and using the bound (19) for $\|\xi^j\|$, this becomes

$$(1+\epsilon)^2\ \|a_{j-1}d^{j-1}\| \le c\ (\epsilon + O(\epsilon^2))\ \|A\|\ (\|x^j\| + \|x^{j-1}\|). \tag{20}$$

It also follows from (8) that

$$a_{j-1}Ap^{j-1} = A(x^j - x^{j-1} - \xi^j),$$

and substituting this expression into the second term in (11) and using the bound (19) for $\|\xi^j\|$, we have

$$(2\epsilon + \epsilon^2)\ \|a_{j-1}Ap^{j-1}\| \le (2\epsilon + O(\epsilon^2))\ \|A\|\ (\|x^j\| + \|x^{j-1}\|). \tag{21}$$

Finally, assume that each term $\|\eta^i\|$, $i = 1, \ldots, j-1$ is bounded by $O(\epsilon)\|A\|(\|x\| + \max_{\ell \le i}\|x^\ell\|)$. It is clear that η^1 satisfies this bound. Since r^{j-1} satisfies

$$r^{j-1} = b - Ax^{j-1} - (b - Ax^0 - r^0) + \sum_{i=1}^{j-1}(A\xi^i + \eta^i),$$

we have, using (14), (16), and the induction hypothesis,

$$\|r^{j-1}\| \le \|A\|\ \|x - x^{j-1}\| + O(\epsilon)\ \|A\|\ (\|x\| + \max_{i \le j-1}\|x^i\|). \tag{22}$$

Substituting (20)–(22) into the bound (11) for $\|\eta^j\|$, we have

$$\begin{aligned}\|\eta^j\| &\le \epsilon\ \|A\|\ \|x - x^{j-1}\| + (2+c)\ \epsilon\ \|A\|\ (\|x^j\| + \|x^{j-1}\|) \\ &\quad + O(\epsilon^2)\ \|A\|\ (\|x\| + \max_{i \le j}\|x^i\|) \\ &\le (\epsilon + O(\epsilon^2))\ \|A\|\ (\|x\| + (5+2c)\max_{i \le j}\|x^i\|). \end{aligned} \tag{23}$$

This shows that $\|\eta^j\|$ is also bounded by $O(\epsilon)\|A\|(\|x\|+\max_{i\le j}\|x^i\|)$, so the induction is complete and (23) is proved. Substituting the bound (23) into (17) and replacing the sum by k times the maximum term gives the desired result. □

Substituting the bounds (14)–(17) into (13) gives the following theorem.

THEOREM 2.2. *The difference between the true residual* $b-Ax^k$ *and the computed vector* r^k *satisfies*

$$\frac{\|b-Ax^k-r^k\|}{\|A\|\ \|x\|} \le (\epsilon+O(\epsilon^2))\ [k+1+(1+c+k\ (10+2c))\ \Theta_k], \tag{24}$$

where c *is defined by* (12) *and* Θ_k *is defined by* (15).

Note that Theorem 2.2 follows from a simple rounding error analysis of formulas (1). No assumptions are made about the coefficients a_{k-1} or the direction vectors p^{k-1} or about whether the algorithm even converges.

The bounds in Lemma 2.1 are not sharp. In particular, if an algorithm is near convergence, then one can expect the norm of r^{j-1} to be much smaller than $\|A\|(\|x\|+\|x^{j-1}\|)$, so the bound on $\|\eta^j\|$ in Lemma 2.1 may be a large overestimate. Still, this bound is roughly the same size as the bound on $\|\xi^j\|$, so if the bound (16) is realistic, then the bound in Theorem 2.2 will be of the right order of magnitude. Based on formula (8) and the rules for floating point arithmetic, one can expect that

$$\|\xi^j\| \sim \epsilon\ \|x^{j-1}\|,$$

so the most significant roundoff error will occur at the step where $\|x^{j-1}\|$ is largest. We can then expect

$$\frac{\|b-Ax^k-r^k\|}{\|A\|\ \|x\|} \ge \epsilon\ \Theta_k.$$

Thus, while the constant terms and the dependence on k in (24) may be overestimates, one can expect the factor Θ_k to play an important role in the size of the difference between the true and computed residuals.

We will demonstrate later that several of the algorithms listed in section 1 generate vectors r^k that approach zero (or at least something much smaller than the machine precision) as $k\to\infty$. It should be noted that once r^{k-1} has been reduced below a certain level, the approximate solution x^k remains essentially unchanged. This is because the norm of the update term $a_{k-1}p^{k-1}$ in (8) is closely related to the size of r^{k-1}, as can be seen from (9),

$$a_{k-1}p^{k-1} = A^{-1}(r^{k-1}-r^k+\eta^k).$$

It follows that if the vectors r^k, and hence $a_{k-1}p^{k-1}$, are converging to zero and if one runs well past the point where $\|A^{-1}r^{k-1}\|$ reaches $O(\epsilon)\|x^{k-1}\|$, the true residual $b-Ax^k$ will not grow like k, as suggested by the bound (24), but will remain almost constant. Denoting by S the number of steps necessary to reach this steady state, the bound (24) can be replaced by

$$\frac{\|b-Ax^k-r^k\|}{\|A\|\ \|x\|} \le (\epsilon+O(\epsilon^2))\ [S+1+(1+c+S\ (10+2c))\ \Theta], \tag{25}$$

where

$$\Theta = \max_j \|x^j\|/\|x\|. \tag{26}$$

Note that if the iteration is started with an extremely large initial guess x^0, say, $\|x^0\| = \epsilon^{-1}$ when $\|x\| = 1$, the factor Θ_k in (24) will be large for all k and, in fact, no method of the form (1) will be likely to find a good approximate solution. This is because even the initial residual cannot be computed accurately, and the coefficients and direction vectors are defined in terms of the updated vectors r^k. We will assume from here on that the initial guess is of reasonable size compared to the true solution and that the term Θ_k becomes large only if the algorithm generates an iterate at some step that is much larger than either the true solution or the initial guess.

3. Specific algorithms. In this section we consider the specific algorithms listed in section 1 and give numerical examples illustrating Theorem 2.2. The numerical tests were performed using MATLAB on a Sparc workstation with machine precision $\epsilon \approx 1.1e-16$. The algorithms were implemented using the formulas given in this section, and no preconditioners were used.

Solid lines in the figures represent actual relative residual norms,

$$\frac{\|b - Ax^k\|}{\|A\|\ \|x\|},$$

while dashed lines show the updated relative residual norms,

$$\frac{\|r^k\|}{\|A\|\ \|x\|}.$$

The value of Θ in (26) is indicated with an asterisk at the step at which the maximum ratio $\|x^j\|/\|x\|$ occurred. In each of the examples the vectors r^k approach zero or something orders of magnitude less than the machine precision as $k \to \infty$. It is therefore expression (25) that determines the ultimately attainable accuracy.

Using exact arithmetic theory, we also derive a priori bounds on the quantity Θ in (26) and argue or demonstrate numerically that the bounds on Θ usually hold in finite precision arithmetic as well. While the result of Theorem 2.2 is independent of any preconditioner, our bounds on Θ may be different when a preconditioner is used, and such differences are noted.

If an algorithm reduces the 2-norm of the error at each step or, more generally, if the iterates satisfy

$$\|x - x^k\| \le \|x - x^0\| \quad \forall k,$$

then we have

$$\|x^k\| \le 2\|x\| + \|x^0\|, \quad \Theta \le 2 + \Theta_0. \tag{27}$$

If the 2-norm of the error can grow, but, say, the B-norm of the error, $\|x - x^k\|_B \equiv \|B^{1/2}(x - x^k)\|$, is reduced at each step, then

$$\|x - x^k\|_B \le \|x - x^0\|_B \Longrightarrow \|x - x^k\| \le \kappa^{1/2}(B)\|x - x^0\|$$

$$\Longrightarrow \|x^k\| \le \|x\| + \kappa^{1/2}(B)(\|x\| + \|x^0\|),$$

and so we have

$$\Theta \leq 1 + \kappa^{1/2}(B)(1 + \Theta_0). \tag{28}$$

To obtain the best a priori bound on the actual residual, we therefore need to determine an error norm that is always reduced over its initial value in the algorithm and that is as close as possible to the 2-norm of the error.

3.1. Steepest descent and the CG method. The steepest descent and CG algorithms for symmetric positive definite problems are as follows.

STEEPEST DESCENT AND CG.

Given an initial guess x^0, compute $r^0 = b - Ax^0$, and set $p^0 = r^0$.
For $k = 1, 2, \ldots,$

Compute $x^k = x^{k-1} + a_{k-1}p^{k-1}$, where $a_{k-1} = \frac{\langle r^{k-1}, r^{k-1}\rangle}{\langle p^{k-1}, Ap^{k-1}\rangle}$.

Set $r^k = r^{k-1} - a_{k-1}Ap^{k-1}$.

Set $p^k = r^k + b_{k-1}p^{k-1}$, where
$b_{k-1} = 0$ for steepest descent, $b_{k-1} = \frac{\langle r^k, r^k\rangle}{\langle r^{k-1}, r^{k-1}\rangle}$ for CG.

It is well known that, in exact arithmetic, each step of these algorithms reduces the A-norm of the error. The steepest descent method minimizes the A-norm of the error at step k in the direction of the residual r^k, while CG minimizes the A-norm of the error over all vectors of the form $P_k(A)e^0$, where e^0 is the initial error and P_k is a kth degree polynomial with value 1 at the origin. It therefore follows from (28) that

$$\Theta \leq 1 + \kappa^{1/2}(A)(1 + \Theta_0), \tag{29}$$

but one can say more. It was shown in [3] for the steepest descent algorithm and in [16] for the CG method that the 2-norm of the error also decreases monotonically. It therefore follows from (27) that

$$\Theta \leq 2 + \Theta_0. \tag{30}$$

Unfortunately, the orthogonality properties used to establish the reduction in 2-norm of the error in CG may fail completely in finite precision arithmetic. An analogy developed in [13] and [14], however, enables one to reach a similar conclusion for finite precision computations. There it was shown that the error in the approximate solution x^k generated by a finite precision CG computation for $Ax = b$ is approximately equal to the error in an approximate solution $\bar{x}^k$ generated by the *exact* algorithm applied to a larger problem $\bar{A}\bar{x} = \bar{b}$, with initial guess $\bar{x}^0$ satisfying $\|\bar{x} - \bar{x}^0\| \approx \|x - x^0\|$. The arguments used to establish monotone convergence of the 2-norm of the error can be applied to the corresponding exact CG iterates $\bar{x}^k$ to obtain

$$\|x - x^k\| \approx \|\bar{x} - \bar{x}^k\| < \|\bar{x} - \bar{x}^0\| \approx \|x - x^0\|.$$

It follows that in finite precision arithmetic, under appropriate assumptions about $\kappa(A)$, one can also expect that (30) will hold.

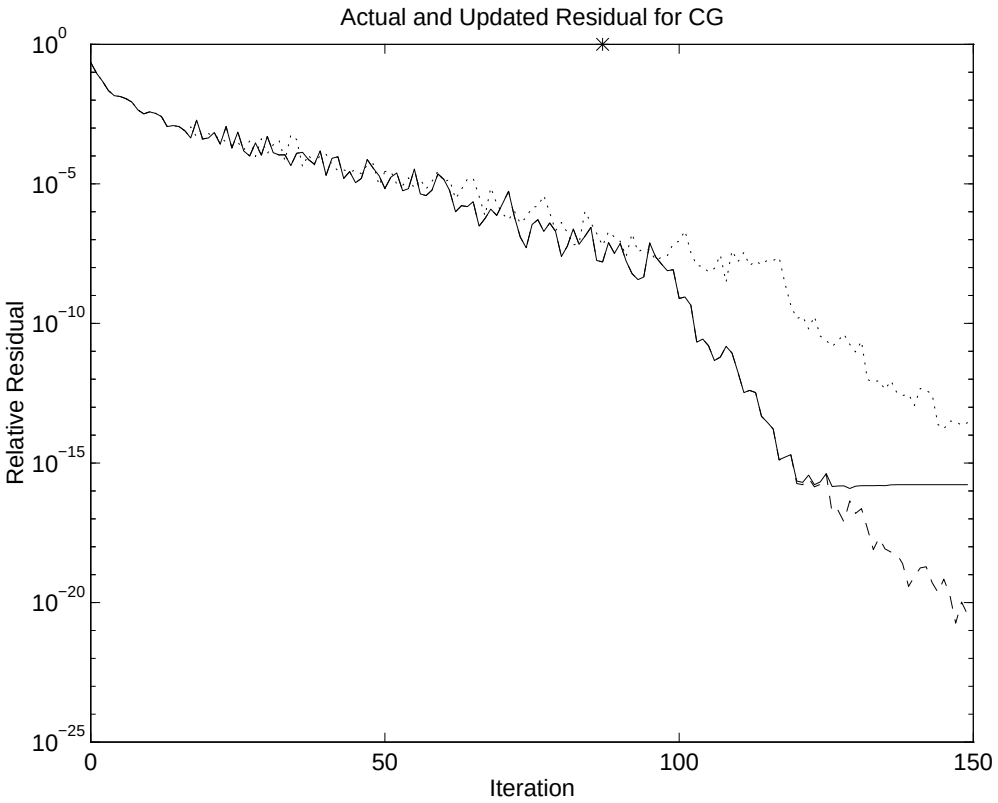


FIG. 1. *Actual residual norm (solid), updated residual norm (dashed), and actual residual norm when residuals are computed directly (dotted).*

We therefore estimate, for both algorithms, that

$$\min_k \frac{\|b - Ax^k\|}{\|A\|\ \|x\|} \leq O(\epsilon)\ S \tag{31}$$

independent of κ. The number of steps S required to reach the steady state for an ill-conditioned problem may be quite large, especially for the steepest descent method, but, as noted previously, the factor S in (25) and, therefore, in (31), is usually an overestimate.

Figure 1 shows the convergence of CG for a problem of size $n = 40$ with eigenvalues geometrically distributed between 1 and 10^4:

$$\lambda_i = \kappa^{(i-1)/(n-1)}, \quad i = 1, \ldots, n, \quad \kappa = 10^4. \tag{32}$$

The matrix A was taken to be of the form $Q\Lambda Q^T$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and Q is a random orthogonal matrix. A random solution was chosen and the initial guess was set to zero.

The solid curve shows the actual relative residual norm, while the dashed curve shows the updated relative residual norm. The two curves coincide until they both

reach the level of machine ϵ, at which point the dashed curve continues to decrease while the solid curve levels off. The asterisk in the figure shows the maximum ratio $\|x^k\|/\|x\|$, plotted at the step at which it occurred. In exact arithmetic, this ratio is bounded by 1 when a zero initial guess is used, and its value in finite precision arithmetic is just slightly larger than 1 for this example. It is clear that rounding errors have a great effect on the rate of convergence of CG for this problem since, in exact arithmetic, the exact solution would be obtained after $n = 40$ steps. Still, the ultimately attainable accuracy is as high as one might expect, with the relative residual reaching the level ϵ and the relative error of size approximately $\kappa(A)\epsilon$. The behavior of the steepest descent method for this problem is not shown, simply because it was so slow to converge. For other less badly conditioned problems, the steepest descent method also attained a final relative residual of size ϵ.

To demonstrate the claim made in section 1—that it is better (in terms of rate of convergence) to update the vectors r^k in the CG algorithm than to compute them directly—we also ran the above algorithm with the formula for r^k replaced by $r^k = b - Ax^k$. The dotted line in the figure shows the residual norm $\|b - Ax^k\|/(\|A\|\,\|x\|)$ for this modified algorithm. This phenomenon of slower convergence was observed before [22], and it has to do with the fact that the normalized residual vectors $(-1)^k r^k/\|r^k\|$ no longer approximately satisfy a three-term recurrence. That is, all of the known proofs of fast convergence for the CG method (convergence at least about as fast as the exact Chebyshev algorithm) [8, 13] rely on the fact that the recurrence can be written in the form

$$AQ_k = Q_k T_k + \beta_k q_{k+1} e_k^T + F_k, \tag{33}$$

where the columns of the n by k matrix Q_k are the normalized residual vectors $r^0/\|r^0\|, -r^1/\|r^1\|, \ldots, (-1)^{k-1} r^{k-1}/\|r^{k-1}\|$, T_k is a k by k tridiagonal matrix $q_{k+1} = (-1)^k r^k/\|r^k\|$, and F_k is a tiny perturbation term; e.g., $\|F_k\| \sim \epsilon\|A\|$. If residuals are computed directly, then the roundoff term F_k is roughly on the order of $\epsilon\|A\|(\|x^k\|/\|r^k\|)$, so when the residual becomes much smaller than the approximate solution, this term is no longer tiny. Of course this does not *prove* that CG will converge slowly if the residuals are computed directly, but it does explain why the known proofs of fast convergence are not applicable.

When a symmetric positive definite preconditioner M is used with either of these algorithms, it is equivalent to applying the unpreconditioned algorithm to the problem $M^{-1/2}AM^{-1/2}y = M^{-1/2}b$, $x = M^{-1/2}y$. The $M^{-1/2}AM^{-1/2}$-norm of $y - y^k$, which is the A-norm of the error, is still minimized at each step, so the bound (29) on Θ still holds. It is possible, however, that the 2-norm of the error $x - x^k$ may grow. We know only that the 2-norm of $y - y^k$, which is the M-norm of $x - x^k$, decreases monotonically. For a preconditioned problem, estimate (31) must therefore be replaced by

$$\min_k \frac{\|b - Ax^k\|}{\|A\|\|x\|} \leq O(\epsilon)\ \min\{\kappa^{1/2}(A), \kappa^{1/2}(M)\}\ S. \tag{34}$$

3.2. The BCG and CGS methods. The BCG algorithm (without look-ahead) for general nonsymmetric problems is as follows.

BCG (WITHOUT LOOK-AHEAD).

Given an initial guess x^0, compute $r^0 = b - Ax^0$, and set $p^0 = r^0$.
Choose $\hat{r}^0$ such that $\langle r^0, \hat{r}^0 \rangle \neq 0$, and set $\hat{p}^0 = \hat{r}^0$. For $k = 1, 2, \ldots$,

Compute $x^k = x^{k-1} + a_{k-1}p^{k-1}$, where $a_{k-1} = \frac{\langle r^{k-1}, \hat{r}^{k-1} \rangle}{\langle Ap^{k-1}, \hat{p}^{k-1} \rangle}$.

Set $r^k = r^{k-1} - a_{k-1}Ap^{k-1}$, $\hat{r}^k = \hat{r}^{k-1} - a_{k-1}A^T\hat{p}^{k-1}$.

Compute $p^k = r^k + b_{k-1}p^{k-1}$, $\hat{p}^k = \hat{r}^k + b_{k-1}\hat{p}^{k-1}$, where $b_{k-1} = \frac{\langle r^k, \hat{r}^k \rangle}{\langle r^{k-1}, \hat{r}^{k-1} \rangle}$.

The BCG algorithm is closely related to the two-sided Lanczos process. For details of this relationship see, for example, [10]. The BCG algorithm breaks down if either

$$\langle Ap^{k-1}, \hat{p}^{k-1} \rangle = 0 \quad \text{or} \quad \langle r^{k-1}, \hat{r}^{k-1} \rangle = 0$$

before an acceptable approximate solution has been obtained. The second condition corresponds to the breakdown of the two-sided Lanczos process, while the first indicates the nonexistence of the BCG iterate. Of course, if either of these quantities is pathologically small, then roundoff is likely to spoil the convergence of the algorithm. In such cases, look-ahead steps can be used to avoid some of the difficulties, at the price of extra work and storage. Even if we assume that these quantities do not become extremely small, however, the norm of the iterate x^k can become arbitrarily large. For this reason, we cannot give an a priori bound on Θ in (26), and, indeed, the algorithm might fail to obtain a small residual even if $\|r^k\| \to 0$.

The CGS algorithm (without look-ahead) for general nonsymmetric problems is as follows.

CGS (WITHOUT LOOK-AHEAD).

Given an initial guess x^0, compute $r^0 = b - Ax^0$, and set $u^0 = r^0$,
$p^0 = r^0$, $q^0 = 0$, and $v^0 = Ap^0$. Choose $\hat{r}^0$ such that $\langle r^0, \hat{r}^0 \rangle \neq 0$.
For $k = 1, 2, \ldots$,

Compute $q^k = u^{k-1} - a_{k-1}v^{k-1}$, where $a_{k-1} = \frac{\langle r^{k-1}, \hat{r}^0 \rangle}{\langle v^{k-1}, \hat{r}^0 \rangle}$.

Compute $x^k = x^{k-1} + a_{k-1}(u^{k-1} + q^k)$.

Set $r^k = r^{k-1} - a_{k-1}A(u^{k-1} + q^k)$.

Compute $u^k = r^k + b_k q^k$, where $b_k = \frac{\langle r^k, \hat{r}^0 \rangle}{\langle r^{k-1}, \hat{r}^0 \rangle}$.

Set $p^k = u^k + b_k(q^k + b_k p^{k-1})$ and $v^k = Ap^k$.

The CGS algorithm requires two matrix–vector multiplications at each step, but no multiplications by the transpose as in BCG. In exact arithmetic, the BCG residual at step k can be written in the form $r_B^k = \phi_k(A)r^0$, where $\phi_k(A)$ is a certain kth

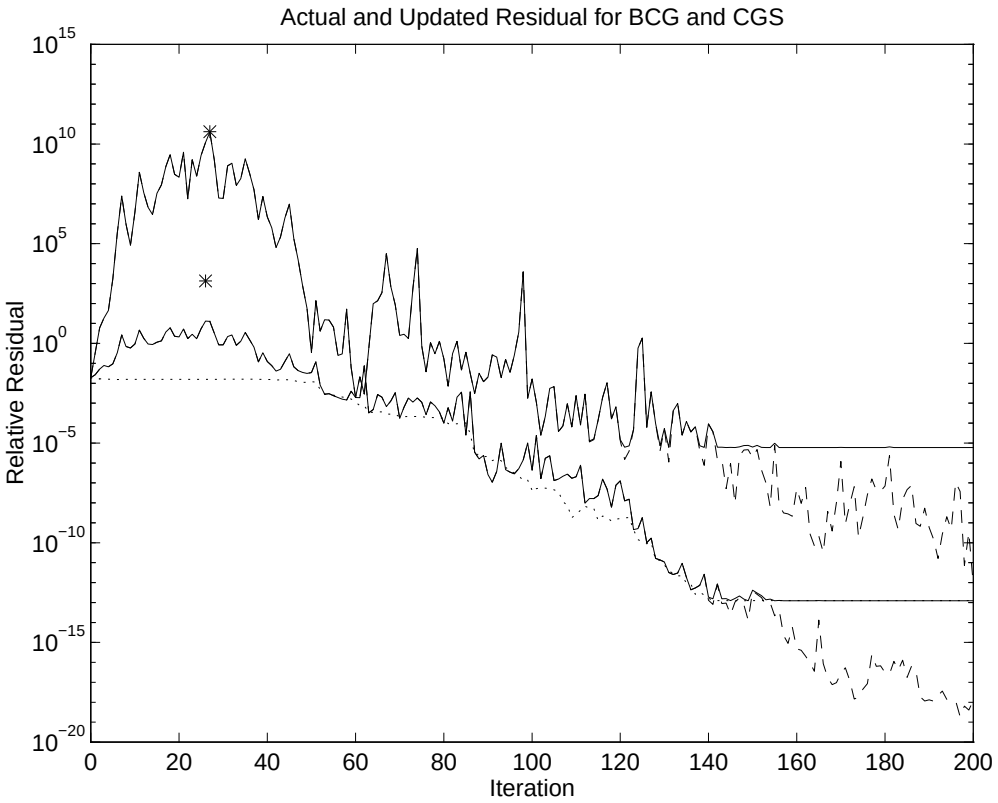


FIG. 2. *Actual residual norm (solid) and updated residual norm (dashed). Top curves are for CGS, bottom ones for BCG. The dotted line shows the actual residual norm for QMR.*

degree polynomial with $\phi_k(0) = 1$. The CGS residual at step k is $r_S^k = \phi_k^2(A)r^0$. If the BCG method is converging well, one would expect the polynomial $\phi_k(A)$ to be small and its square to be even smaller. If the norm of $\phi_k(A)$ is large, however, then that of $\phi_k^2(A)$ will be even larger. Again, the norms of the iterates x^k can become arbitrarily large and we cannot give an a priori bound on Θ in (26).

An example is shown in Figure 2. Here A was taken to be a discretization of the convection diffusion operator

$$-\triangle u + 40(xu_x + yu_y) - 100u \tag{35}$$

on the unit square with Dirichlet boundary conditions, using centered differences on a 32×32 mesh. Similar problems were considered in [11] and [4]. The solution was taken to be $u(x, y) = x(x-1)^2y^2(y-1)^2$ and the initial guess was set to zero. The initial vector $\hat{r}^0$ was set equal to r^0.

The lower solid line in the figure represents the true BCG residual norm, while the lower dashed line shows the updated residual norm. The lower asterisk indicates that the maximum ratio $\|x^k\|/\|x\|$ was approximately 10^3 for BCG. The final actual residual is of size $1.e-13 \approx 10^3\epsilon$, as predicted. Note that in this example the relative residual norm for BCG grows only to about 10^1, so the size of the final residual cannot be predicted from the size of intermediate residuals, but only from the size of intermediate iterates.

The upper solid and dashed lines in Figure 2 show the true and updated residual norms for CGS. The higher asterisk indicates that for CGS the norm of an intermediate iterate grew to approximately $4 \cdot 10^{10}$ times the norm of the true solution, and so, as predicted, the final actual residual reaches the level $6.e-6$, which is roughly $4 \cdot 10^{10}\epsilon$, instead of ϵ. Note that the example presented here was chosen specifically to illustrate the possibility of large growth in iterates and the corresponding loss of accuracy. It should not necessarily be interpreted as showing the typical behavior of the CGS algorithm.

For comparison with BCG, the dotted line in the figure shows the true QMR residual norm for the same problem when the implementation of [2] is used. The QMR convergence curve looks very much like the lower envelope of the BCG convergence curve (as shown in [7] for exact arithmetic), and QMR achieves about the same size final residual as BCG. This cannot be explained using the analysis of this paper, however, because QMR is not of the form (1) and the updated QMR residual vector (not shown) does not converge to zero but levels out at a slightly smaller value than the true QMR residual norm.

3.3. CGNE and CGNR. The CGNE and CGNR algorithms for general nonsymmetric problems are as follows.

CGNE AND CGNR.

Given an initial guess x^0, compute $r^0 = b - Ax^0$, and set $p^0 = A^T r^0$.
For $k = 1, 2, \ldots,$

Compute $x^k = x^{k-1} + a_{k-1}p^{k-1}$, where
$a_{k-1} = \frac{\langle r^{k-1}, r^{k-1}\rangle}{\langle p^{k-1}, p^{k-1}\rangle}$ for CGNE, $a_{k-1} = \frac{\langle A^T r^{k-1}, A^T r^{k-1}\rangle}{\langle Ap^{k-1}, Ap^{k-1}\rangle}$ for CGNR.

Set $r^k = r^{k-1} - a_{k-1}Ap^{k-1}$.

Compute $p^k = A^T r^k + b_{k-1}p^{k-1}$, where
$b_{k-1} = \frac{\langle r^k, r^k\rangle}{\langle r^{k-1}, r^{k-1}\rangle}$ for CGNE, $b_{k-1} = \frac{\langle A^T r^k, A^T r^k\rangle}{\langle A^T r^{k-1}, A^T r^{k-1}\rangle}$ for CGNR.

In exact arithmetic, CGNE is equivalent to the CG algorithm for the linear system $AA^T y = b$, $x = A^T y$. If y^k and $\hat{p}^k$ are the iterate and direction vector at step k of the CG algorithm for this problem, then $x^k = A^T y^k$ and $p^k = A^T \hat{p}^k$. Since CG applied to $AA^T y = b$ minimizes the AA^T-norm of $y - y^k$, CGNE minimizes the equivalent quantity, the 2-norm of the error $x - x^k$. It follows that, in exact arithmetic, $\Theta \leq 2 + \Theta_0$, and since this estimate does not require the orthogonality of previous residual vectors, it can be expected to hold in finite precision arithmetic as well.

In exact arithmetic, CGNR is equivalent to the CG algorithm for the linear system $A^T Ax = A^T b$. The iterates and direction vectors are the same as for the CG algorithm applied to the normal equations, but the vectors r^k are the residuals of the original linear system $b - Ax^k$. Since CG applied to $A^T Ax = A^T b$ minimizes the $A^T A$-norm of $x - x^k$, CGNR minimizes the equivalent quantity, the 2-norm of the residual $b - Ax^k$. Since the 2-norm of the residual is reduced at each step, it follows from (28) that $\Theta \leq 1 + \kappa(A)(1 + \Theta_0)$. Since the 2-norm of the error is also reduced at each step by the CG algorithm applied to $A^T Ax = A^T b$, however, it follows that the same holds for CGNR. As argued previously, the CG method can really be expected to reduce

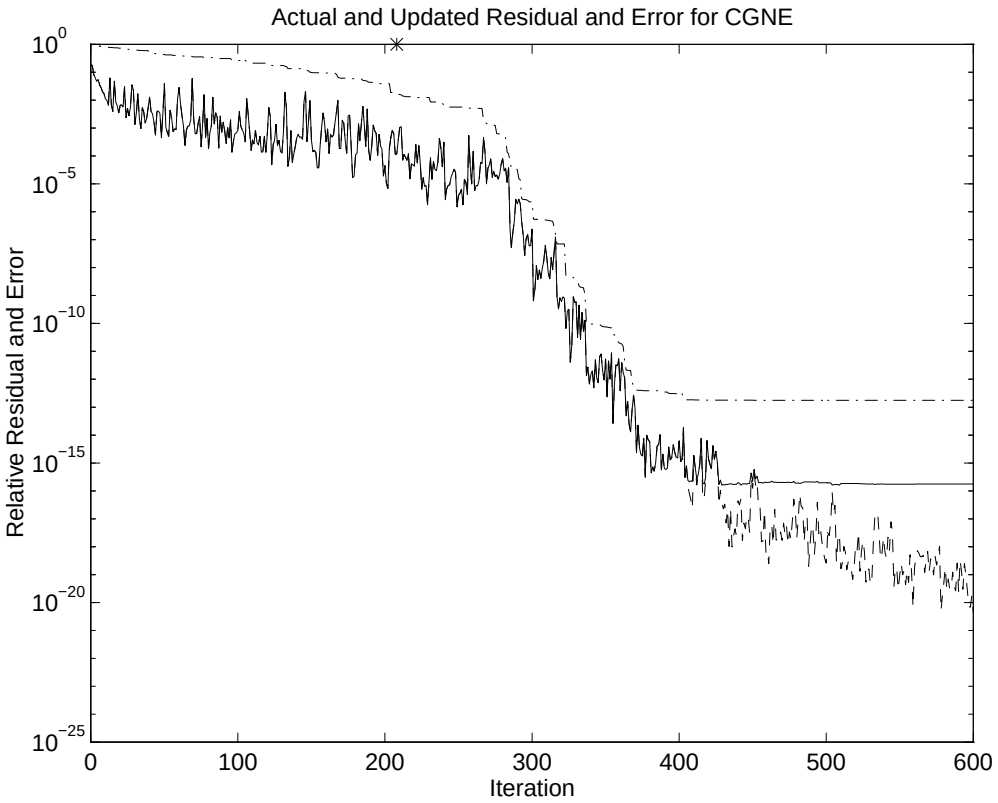


FIG. 3. *Actual residual norm (solid), updated residual norm (dashed), and actual error norm (dash–dot).*

the 2-norm of the error in finite precision arithmetic, based on the analogy developed in [13] and [14]. It therefore follows from (27) that $\Theta \leq 2 + \Theta_0$.

For both algorithms, we therefore estimate that

$$\min_k \frac{\|b - Ax^k\|}{\|A\| \ \|x\|} \leq O(\epsilon) \ S \tag{36}$$

independent of κ. Since this implies that the relative error, $\min_k \|x - x^k\|/\|x\|$, is bounded by $\kappa(A) \ O(\epsilon) \ S$, there is no loss of accuracy associated with the squared condition number of the normal equations (except for what might result from a possibly larger value of S). There may be tighter restrictions on κ, necessary to ensure the convergence of the vectors r^k to zero, but if these restrictions are met, then the accuracy attainable from CGNE and CGNR is as great as that attainable from CG.

Figure 3 shows the actual and updated residual norms as well as the actual error norm, $\|x - x^k\|/\|x\|$, for CGNE applied to a problem with singular values geometrically distributed between 1 and 10^4. The matrix A was taken to be of the form $A = U\Sigma V^T$, where $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_n)$, with $\lambda_1, \ldots, \lambda_n$ defined by (32), and U and V random orthogonal matrices. The initial guess was zero and the solution was random. As can be seen from the figure, the relative residual norm reaches the level ϵ, while the relative error norm reaches the level $\kappa(A)\epsilon$. The maximum ratio Θ is approximately 1, as indicated by the asterisk in the figure. Note that the jumps in the residual norm

do not cause any loss of accuracy since the iterate norms remain bounded by 1. The actual and updated residual norm curves for CGNR applied to the same problem (not shown) look like smoothed-out versions of those for CGNE. The CGNR algorithm also obtained a final relative residual of size ϵ and a final relative error of size $\kappa(A)\epsilon$.

Similar tests were performed for matrices with singular values geometrically distributed between 1 and $10^1, 10^2, \ldots, 10^8$. Although some of these problems required many thousands of steps to reach the steady state, in all cases the actual relative residual norm for both algorithms leveled out at a value less than 10ϵ.

Since preconditioned CGNE is equivalent to CG applied to the normal equations $AA^Ty = b$, with a preconditioner MM^T, the algorithm still minimizes the AA^T-norm of $y - y^k$, which is the 2-norm of $x - x^k$. Estimate (36) therefore holds for the preconditioned algorithm as well.

When CGNR is preconditioned with a matrix M, it is equivalent to applying CG to the normal equations $A^TAx = A^Tb$ using the preconditioner M^TM. This is equivalent to applying unpreconditioned CG to the system $M^{-T}A^TAM^{-1}y = M^{-T}A^Tb$, $x = M^{-1}y$. The $M^{-T}A^TAM^{-1}$-norm of $y - y^k$, which is the A^TA-norm of $x - x^k$, is minimized over a Krylov space, and so it follows from (28) that $\Theta \leq 1+\kappa(A)(1+\Theta_0)$. Since the 2-norm of $y - y^k$, which is the M^TM-norm of $x - x^k$, also decreases monotonically, it also follows from (28) that $\Theta \leq 1 + \kappa(M)(1 + \Theta_0)$. The bound (36) must therefore be replaced by

$$\min_k \frac{\|b - Ax^k\|}{\|A\|\|x\|} \leq O(\epsilon)\ \min\{\kappa(A), \kappa(M)\}\ S \tag{37}$$

for preconditioned CGNR.

4. Alternate implementations. From the preceding discussion it is clear that if one wishes to use an iterative method in which the norms of intermediate iterates can grow without bound, then one should avoid the use of the update formulas in (1) if the iterate norms become too large. A number of algorithms have been developed to do this. For symmetric indefinite problems, the SYMMLQ method of Paige and Saunders [17] accomplishes this goal by updating certain well-determined intermediate quantities instead of the possibly ill-determined iterates x^k. The composite step biconjugate gradient method of Bank and Chan [1] avoids large iterates by skipping steps at which they appear.

The other possibility for avoiding large iterates is to use methods of the form (1) that reduce some error norm that is not too different from the 2-norm. The ORTHOMIN and ORTHODIR methods, for example, minimize the 2-norm of the residual at each step, so that *if* they were implemented in the form (1) (which would require an extra matrix–vector multiplication at each step) then one would expect to obtain a final relative residual of size $\kappa(A)\ O(\epsilon)\ S$ in cases where the updated vectors r^k converge to zero. The QMR method minimizes the norm of a quantity that differs from that of the residual by no more than a factor $\sqrt{k+1}$ at step k, so that *if* QMR were implemented in the form (1) and assuming that the updated vectors r^k became much smaller than the machine precision, one would obtain a final residual of size less than $\kappa(A)\ O(\epsilon)\ S^{3/2}$.

Finally, while it is important to be aware of the ultimately attainable accuracy of an iterative method, this is often less important in practice than the convergence rate of the method before the ultimately attainable accuracy is achieved. Unfortunately, the analysis of convergence rates in finite precision arithmetic (and sometimes even in exact arithmetic) may be a much more difficult problem.

REFERENCES

[1] R. E. Bank and T. F. Chan, *A composite step bi-conjugate gradient algorithm for solving nonsymmetric systems*, Numer. Algorithms, 7 (1994), pp. 1–16 .

[2] R. Barrett, M. Berry, T. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. van der Vorst, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM, Philadelphia, PA, 1993.

[3] J. A. M. Bollen, *Numerical stability of descent methods for solving linear equations*, Numer. Math., 43 (1984), pp. 361–377.

[4] T. F. Chan and T. Szeto, *The composite step family of nonsymmetric conjugate gradient methods*, in PCG '94: Advances in Numerical Methods for Large Sparse Sets of Linear Equations, M. Natori and T. Nodera, eds., Keio University, Yokohama, Japan, 1994, pp. 215–228.

[5] C. Concus, G. H. Golub, and D. P. O'Leary, *A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations*, in Sparse Matrix Computations, J. R. Bunch and D. J. Rose, eds., Academic Press, New York, 1976, pp. 309–332.

[6] E. J. Craig, *The N-step iteration procedures*, J. Math. Phys., 34 (1955), pp. 64–73.

[7] J. Cullum and A. Greenbaum, *Relations between Galerkin and norm-minimizing iterative methods for solving linear systems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 223–247.

[8] V. Druskin and L. Knizhnerman, *Error bounds in the simple Lanczos procedure for computing functions of symmetric matrices and eigenvalues*, Comput. Math. Math. Phys., 31 (1991), pp. 20–30.

[9] R. Fletcher, *Conjugate gradient methods for indefinite systems*, in Numerical Analysis Dundee 1975, G. A. Watson, ed., Lecture Notes in Mathematics 506, Springer-Verlag, Berlin, 1976, pp. 73–89.

[10] R. W. Freund, G. H. Golub, and N. M. Nachtigal, *Iterative solution of linear systems*, Acta Numerica, 1 (1992), pp. 57–100.

[11] R. W. Freund and N. M. Nachtigal, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.

[12] R. W. Freund and N. M. Nachtigal, *An implementation of the QMR method based on coupled two-term recurrences*, SIAM J. Sci. Comput., 15 (1994), pp. 313–337.

[13] A. Greenbaum, *Behavior of slightly perturbed Lanczos and conjugate gradient recurrences*, Linear Algebra Appl., 113 (1989), pp. 7–63.

[14] A. Greenbaum and Z. Strakos, *Predicting the behavior of finite precision Lanczos and conjugate gradient computations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 121–137.

[15] N. J. Higham and P. A. Knight, *Componentwise error analysis for stationary iterative methods*, in Linear Algebra, Markov Chains, and Queueing Models, C. D. Meyer and R. J. Plemmons, eds., IMA Volumes in Mathematics and its Applications 48, Springer-Verlag, New York, 1993, pp. 29–46.

[16] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 409–435.

[17] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 11 (1974), pp. 197–209.

[18] P. Sonneveld, *CGS, a fast Lanczos-type solver for nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 36–52.

[19] H. A. van der Vorst, *Bi-CGSTAB: A fast and smoothly converging variant of bi-CG for the solution of nonsymmetric linear systems*, SIAM J. Sci. Comput., 13 (1992), pp. 631–644.

[20] P. K. W. Vinsome, *Orthomin, an iterative method for solving sparse sets of linear equations*, in Proc. Fourth Symposium of Reservoir Simulation, Society of Petroleum Engineers of AIME, 1976, pp. 149–159.

[21] H. Wozniakowski, *Round-off error analysis of iterations for large linear systems*, Numer. Math., 30 (1978), pp. 301–314.

[22] H. Wozniakowski, *Round-off error analysis of a new class of conjugate gradient algorithms*, Linear Algebra Appl., 29 (1980), pp. 507–529.

[23] D. M. Young and K. C. Jea, *Generalized conjugate gradient acceleration of nonsymmetrizable iterative methods*, Linear Algebra Appl., 34 (1980), pp. 159–194.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 3, pp. 552–565, July 1997

003

FAST NESTED DISSECTION FOR FINITE ELEMENT MESHES *

SHANG-HUA TENG†

Abstract. We present a randomized $O(n \log\log n)$ time algorithm for constructing a recursive separator decomposition for well-shaped meshes in two and three dimensions. Our algorithm takes $O(n \log\log n)$ time while previous algorithms require $\Theta(n \log n)$ time. It uses techniques from probability theory, computational geometry, and graph theory. The new algorithm has an application in the solution of sparse linear systems that arise in finite element calculations. In particular, it can be used to design $O(n \log\log n)$ time algorithms for finding a provably good nested-dissection ordering for 3D finite element systems. It can also be used to improve the construction of 3D point location structures, which are useful in hierarchical methods such as the multigrid and multilevel domain decompositions.

Key words. finite element meshes, geometric sampling, hierarchical computing, the multigrid, multilevel methods, nested dissection, 3D point location, separators, sparse matrix computations

AMS subject classifications. 05C50, 65F50, 68R10

PII. S0895479895284282

1. Introduction.

Large sparse linear systems arise in many areas of scientific applications. The solution to these linear systems is a key step in commonly used methods such as finite element and finite difference methods [33]. Many practical applications, such as circuit design, require an exact solution to a linear system and hence a direct method that uses Gaussian elimination is desired. Direct methods are also used as a subroutine in preconditioning-based iterative methods.

Gaussian elimination may introduce fill in solving sparse linear systems. The amount of fill depends on the ordering of the rows and columns of the sparse matrix. Such an ordering is called an *elimination ordering*. The problem of finding an optimal ordering is known to be NP-hard. *Nested dissection* [1, 12, 13, 16, 26] is one of the provably good methods for finding an elimination ordering for sparse matrices. The nested-dissection algorithm divides the graph of a linear system by first finding a small *separator*, that is, a set of vertices whose removal divides the rest of the graph into two disjoint subgraphs of approximately equal size, and then recursively divides the two subgraphs. The vertices in the two subgraphs are ordered recursively and are placed ahead of those in the separator. In this procedure, we do not require the two subgraphs to be connected themselves, for if one of the subgraphs has several connected components, then we can order each component independently. For 3D applications, the *geometric separator algorithm* of Miller et al. [30, 31] can be used to perform nested dissection [17]. However, the geometric separator algorithm or

* Received by the editors April 7, 1995; accepted for publication (in revised form) by J. W. H. Liu July 8, 1996. This research was supported in part by an NSF CAREER award (CCR-9502540) and an Alfred P. Sloan Research Fellowship. Part of this work was done while the author was at the Department of Mathematics and the Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA, where his work was supported in part by AFOSR F49620-92-J-0125 and Darpa N00014-92-J-1799.

http://www.siam.org/journals/simax/18-3/28428.html

† Department of Computer Science, University of Minnesota, Minneapolis, MN 55455 (steng@cs.umn.edu).

any other separator algorithm takes at least linear time to find a small separator; nested-dissection methods[1] use at least $\Omega(n \log n)$ time. The application of Miller et al.'s geometric separator algorithm in three dimensions achieves the $O(n \log n)$ time bound. In this paper, we give a randomized $O(n \log \log n)$ time algorithm.

The solution to a sparse linear system usually involves four steps:

1. find a good elimination ordering,
2. perform a symbolic factorization to predict the amount of fill and create a data structure with the proper amount of space for the numerical factorization,
3. compute the numerical factorization which expresses the reordered sparse matrix as a product of two triangular systems,
4. solve the resulting triangular linear systems.

Our result improves the time needed for the ordering by a factor of $O(\log n / \log \log n)$.

Algorithmically, we show how to construct a *recursive separator decomposition* for well-shaped 3D meshes in $O(n \log \log n)$ time. Our construction uses *sampling* to reduce the time needed for finding good separators. It then applies a *geometric data structure* to assist a graph connectivity algorithm to speed up the recursive separator decomposition. Our algorithm also improves the 3D point-location algorithm of [32, 36], which has potential applications to hierarchical computing.

In section 2, we review the graph partitioning problem and well-shaped finite element meshes. In section 3, we review the geometric separator algorithm of Miller et al. In section 4, we present and analyze our $O(n \log \log n)$ time recursive separator decomposition algorithm. In section 5, we discuss its applications in nested dissection and 3D point location for finite element meshes. In section 6, we discuss some issues in the potential implementation of techniques developed in this paper. We give a high-level outline of a geometric sampling-based multilevel algorithm for partitioning and ordering. The proofs presented in section 4 can be extended to show that this multilevel algorithm generates provably good multiway partitions and nested-dissection orderings.

2. Separators and well-shaped meshes. In this paper, we will use the following definition of vertex separators.

DEFINITION 2.1 (separators). *A subset of vertices C of a graph G with n vertices is an* $f(n)$-separator *that* δ-splits *if $|C| \leq f(n)$ and the vertices of $G - C$ can be partitioned into two sets A and B such that there are no edges from A to B and $|A|$, $|B| \leq \delta n$, where f is a function and $0 < \delta < 1$.*

In this definition and for the rest of the paper, $|A|$ denotes the cardinality of finite set A. The type of separator defined here is sometimes called a "vertex separator," that is, a subset C of vertices of G whose removal disconnects the graph into two or more graphs of smaller size. A related concept is an "edge separator," that is, a set of edges whose removal disconnects the graph.

A *separator tree* is a tree structure generated by the recursive applications of a separator algorithm. The root of the tree corresponds to the top-level separator. The root has two children for the two subgraphs induced by the top-level separator, respectively. Notice again that we do not require the two subgraphs to be connected themselves. If one of them has more than one component, we still keep all of its components together. Therefore, the "subgraph" that we use may not be connected. So the input graph to our graph partitioning algorithm and in Definition 2.1 need not

[1] Recently, Goodrich [19] gave a linear time algorithm for finding a recursive separator decomposition of a planar graph. However, instead of using $O(n^{0.5})$-separators, Goodrich used $O(n^{0.5+\epsilon})$-separators.

be connected.

The subtrees of the root are recursively generated for the two subgraphs. Associated with each internal node of the separator tree is a subgraph that is induced by the set of separators used in the path from the root to the internal node. Clearly, the graph associated with the root is the original graph. We call a separator tree an $f(n)$-*separator tree* if it is generated by applying an $f(m)$-separator to each internal node of the tree whose subgraph has size m.

In this paper we consider the graph associated with a well-shaped unstructured mesh that arises from finite element calculations.

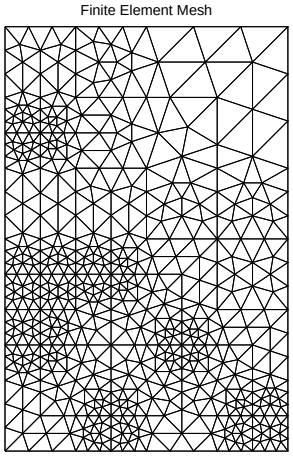


FIG. 1. *A well-shaped mesh in two dimensions (courtesy of David Eppstein).*

A *mesh* is a decomposition of a domain into a collection of simple elements (see Fig. 1 for a 2D example). A common choice for an element in the finite element method is a d-dimensional *simplex*, e.g., a triangle in two dimensions and a tetrahedron in three dimensions. A d-dimensional *simplicial complex* is defined to be a collection of d-dimensional simplices that meet only at shared faces [4, 5]. So a 2D simplicial complex is a collection of triangles that intersect only at shared edges and vertices. The "corners" of the simplices of a mesh are called the *vertices* of the mesh. Associated with each mesh is a natural graph, its 1-*skeleton*, which is a graph defined on the vertices of the mesh where there is an edge between two vertices iff they are both contained in a simplex of the mesh. The finite element formulation on a mesh defines a sparse matrix, called the *stiffness matrix* of the mesh (see the paragraph below). The nonzero pattern of the stiffness matrix is the 1-skeleton of the mesh.

A mesh is given as a list of its elements, where each element is given by the information describing the hierarchical structure of the element, i.e., its lower-dimensional structures that include its faces, edges, and vertices. Moreover, each vertex has geometric coordinates in two or three dimensions. In the finite element method, a linear system is defined over a mesh, with variables representing physical quantities at the nodes. To properly approximate a continuous function, in addition to the conditions that a mesh must conform to the boundary of the region and be fine enough, each individual element of the mesh must be *well shaped*. A common shape criterion for elements is the condition that the angles of each element are not too small or the aspect ratio of each element is bounded [5, 10, 33]. We refer the reader to [30] for a detailed discussion of these geometric conditions. We call meshes that have these

geometric conditions *well-shaped meshes*. We now summarize some useful properties of well-shaped meshes.

- The degree of each vertex of a well-shaped mesh is bounded by a constant, where the *degree* of a vertex is the number of edges in the 1-skeleton incident to the vertex. Hence, the mesh has a linear number of edges.
- Each well-shaped mesh of n vertices has an $O(n^{1-1/d})$-separator and satisfies Theorem 3.1 in the next section.

We will use the following geometric characterization of a well-shaped mesh. Suppose $M = (A, xyz)$ is the input mesh, where A describes the combinatorial structure of the mesh: its nodes V, edges E, and elements F. For each point $p \in xyz$, let B_p be the ball centered at p with a radius that is equal to the distance from p to the nearest point q in xyz such that (p, q) is an edge of M. We call $\Gamma = \{B_p : p \in xyz\}$ the *nearest neighborhood system* of xyz. Because the mesh has a linear number of edges, we can construct its nearest neighborhood system in linear time.

THEOREM 2.2 (see [31]). *Let $M = (A, xyz)$ be a well-shaped mesh in $\mathbb{R}^d$. Then there exists a constant $\alpha > 1$, depending only on d and the aspect ratio of elements in M, such that for each edge (u, v) of M, B_u intersects $\alpha \cdot B_v$ and $\alpha \cdot B_u$ intersects B_v, where $\alpha \cdot B_v$ is the ball with the same center as B_v but with a radius that is α times the radius of B_v.*

3. Geometry mesh partitioning: A review. Miller et al. [30] showed that the geometric structure of a well-shaped mesh can be used to develop a quality guaranteed separator algorithm. Their geometric separator algorithm divides a mesh using a sphere. Suppose (A, xyz) is the input mesh. Recall that A describes the combinatorial structure of the mesh: its nodes V, edges E, and elements F. The set $xyz \in \mathbb{R}^d$ gives the geometric coordinates of the mesh (in d dimensions). The geometric separator algorithm finds a $(d-1)$-sphere S in $\mathbb{R}^d$ that divides nodes, edges, and elements of the mesh into three subsets: those nodes V_I, edges E_I, and elements F_I that are in the interior of S; those V_E, E_E, and F_E that are in the exterior of S; and those V_O, E_O, and F_O that form a vertex-, edge-, and element-separator, respectively, induced by S. In the notation above, the subscripts I and E stand for "interior" and "exterior," respectively, and the subscript O stands for "overlap with the sphere." In this paper, a $(d-1)$-sphere is the boundary of a d-dimensional ball. The details of the theory and implementation of the geometric separator algorithm can be found in [18, 30]. We now give a high-level description of the algorithm, which is necessary for our discussions in the remainder of the paper.

ALGORITHM (geometric partition).
Input A well-shaped mesh (A, xyz) in $\mathbb{R}^d$.
Output A $(d-1)$-sphere S in $\mathbb{R}^d$.

1. Choose a set P of random samples (of about 1000 points) from xyz;
2. Compute the stereo graphic image Q on U_d of P, where U_d is a unit d-sphere in $\mathbb{R}^{d+1}$;
3. Find an approximate centerpoint c of Q;
4. Conformally map c to the sphere center of U_d (see [30] for the formula). The mapping can be represented as a $(d+1) \times (d+1)$ matrix and can be found in $O(d)$ time.
5. Choose a random great circle C of U_d;
6. Transform C back to $\mathbb{R}^d$ to obtain S and compute the vertex quality of S.

The *vertex quality* of a separating sphere S is the size of vertex-separator V_O induced by S. There are several ways to derive a vertex-separator V_O from S [30]. We will use the following approach. Let $\Gamma = \{B_p : p \in xyz\}$ be the nearest neighborhood system of xyz as defined in section 2. Let V_O be the set of the vertices whose balls either (i) intersect S or (ii) are smaller than S and would intersect S if magnified by a factor of α, where α is the constant given in Theorem 2.2. Therefore, V_E will be the set of vertices not in V_O that are located in the exterior of S, and V_I will be the set of vertices not in V_O that are located in the interior of S. We now show that no vertex in V_I is connected with any vertex in V_E in M. For each pair of points $u \in V_E$ and $v \in V_I$, because neither u nor v is in V_O, we conclude that B_u and B_v do not intersect S. Because v is in the interior of S, B_v must be smaller than S (for otherwise B_v would intersect S). Because v does not belong to V_O, $\alpha \cdot B_v$ does not intersect S either. Thus, B_u does not intersect $\alpha \cdot B_v$. It then follows from Theorem 2.2 that (u, v) cannot be an edge of M. Therefore, V_O is a vertex-separator that divides V_E from V_I.

Computationally, there are two major steps in the geometric separator algorithm: (1) centerpoint computation and the selection of S and (2) testing the quality of the returned sphere. Step 1 takes constant time. Given the nearest neighborhood system Γ, we can determine in constant time whether a vertex is in V_O. The time to construct V_O is thus linear in n. Step 2 takes linear time. We will repeat the above procedure until the returned separator is in the order of $O(n^{1-1/d})$. The following theorem summarizes the properties of the geometric separator algorithm that will be used in the next section.

THEOREM 3.1 (see [31]). *For each well-shaped mesh $M = (A, xyz)$, there exists a constant δ in the range of $(d+1)/(d+2) \le \delta < 1$, such that the geometric separator algorithm finds a $(d-1)$-sphere S such that* (1) $|V_I|, |V_E| \le \delta|V|$ *and* $|F_I|, |F_E| \le \delta|F|$, (2) $|V_O| = O(n^{1-1/d})$, $|E_O| = O(n^{1-1/d})$, *and* $|F_O| = O(n^{1-1/d})$. *Moreover, with probability at least* $1/2$, *the returned sphere S satisfies these quality conditions.*

4. Fast separator tree decomposition. Even though the algorithm given in section 3 finds a sphere S in constant time and the probability that S is good (namely, $|V_O| = O(n^{1-1/d})$) is at least $1/2$, we still need $\Theta(n)$ additional time to compute the vertex quality of S to make sure that we can use S as the root-separator of the tree. Thus, the naive recursion of the geometric separator algorithm requires $\Omega(n \log n)$ time to build a separator tree. In this section, we will reduce the time to $O(n \log \log n)$.

4.1. Representation of a separator tree. In the context of this paper, a vertex-separator of a mesh is derived from a sphere. Every sphere is given by its center and radius, so we only need a constant amount of space to store the information of a sphere.

Given a mesh $M = (A, xyz)$, the recursive applications of the geometric separator algorithm generate a *tree of spheres*. The root of the tree corresponds to the top-level sphere S, which induces a vertex-separator V_O that divides the mesh into two submeshes M_E and M_I, where M_E and M_I are submeshes defined by V_E and V_I, respectively. The root has two subtrees which are generated for M_I and M_E, respectively: its left subtree is for M_I and its right subtree is for M_E. In this "tree of spheres" representation, we do not store vertex-separators; we only store spheres. The separator tree defined in section 2 stores a vertex-separator at each internal node.

In our algorithm below, we will first construct a tree of spheres representation of an $O(n^{1-1/d})$-separator tree. We will then compute the vertex-separator of each internal node. Note that it takes constant time for the geometric separator algorithm

to return a sphere, but linear time to compute the vertex-separator induced by the sphere. Our objective is to use this fact to build the tree of spheres representation faster than $O(n \log n)$ time. However, to guarantee the quality of the separator tree, we need to estimate the size of the vertex-separator induced by the sphere. We need to perform this estimation faster than linear time. We will estimate the size of the vertex-separator by sampling.

4.2. A high-level discussion. The use of sampling is to reduce the time needed at each level of the separator tree. This idea can be depicted by the following high-level construction: choose a sample Γ_1 of $O(n/\log n)$ balls uniformly at random from the nearest neighborhood system $\Gamma = \{B_p : p \in xyz\}$ of the input mesh. Construct a separator tree T_1 for Γ_1. We will stop the recursion when the number of balls is $O((\log n)^h)$ for a constant h to be specified later. At each internal node of T_1, we compute the vertex-separator of the sphere based on Γ_1 (instead of Γ). Since we deal only with $O(n/\log n)$ balls, the total complexity is at most $O(n)$ for the construction of the tree of spheres representation of T_1. In the next subsection, we will show that with high probability T_1 is the initial fraction of an $O(n^{1-1/d})$-separator tree for the input mesh. Moreover, the size of the submesh associated with each leaf of T_1 is $\Theta((\log n)^h)$.

Our next step is to "identify" the vertex-separator of the input mesh induced by the sphere at each internal node of T_1 and the submesh associated with every leaf of T_1. Because the size of the submesh at each leaf of T_1 is $\Theta((\log n)^h)$, T_1 has $\Theta(n/(\log n)^h)$ leaves. We can then apply the standard nested-dissection method to all leaves of T_1 to complete the separator tree construction. For each leaf, we need $O((\log n)^h \log\log n)$ time. Therefore, we can complete the separator tree from T_1 in $O(n \log\log n)$ time.

However, the identification problem is nontrivial. We will present a solution that uses a geometric search structure to simulate a graph connectivity algorithm to obtain the following algorithmic result.

THEOREM 4.1 (main). *Given a well-shaped mesh $M = (A, xyz)$, we can find an $O(n^{1-1/d})$-separator tree decomposition of (A, xyz) in random $O(n \log\log n)$ time with high probability.*

4.3. Sampling for the construction of a partial separator tree. Notice that the geometric separator algorithm takes only constant time to find a top-level sphere separator (because finding an approximate centerpoint uses only constant time [8, 35]). But to build an $O(n^{1-1/d})$-separator tree, we need to test the quality of the sphere. It takes $O(n)$ time to do so deterministically. Our idea to reduce the quality-testing time is to use sampling to approximate the separator size: *To approximately count the size of the vertex-separator, we choose a sample of $n/\log n$ vertices and count how many of them are in the separator.* The quality-testing time is reduced to $O(n/\log n)$. The following classical result from probability theory guarantees the quality and correctness of our sampling method.

LEMMA 4.2 (Chernoff–Hoeffding). *There is a constant $c > 1$ such that the following is true: Suppose there are L red balls in a set of n balls. Then for all $s(n)$, if we take a sample of $s(n)$ random balls from the set and observe r red balls,*

$$\mathbf{Prob}[r/(2s(n)) \le L/n \le 2r/s(n)] \ge 1 - e^{-cs(n)L/n}.$$

Proof. This lemma is a special case of Chernoff–Hoeffding bounds, which bound the probability that a sum of independent random 0-1 variables differs significantly

from its mean [2, 20, 22]. Suppose $S = \sum_{i=1}^{m} X_i$, where X_i is a random 0-1 variable from an appropriate distribution. The most commonly used form of the Hoeffding bounds states that for any $0 \leq \epsilon \leq 1$ the following holds:

$$\textbf{Prob}[S \leq (1-\epsilon)E[S]] \leq e^{-\epsilon^2 E[S]/3}, \tag{1}$$

$$\textbf{Prob}[S \geq (1+\epsilon)E[S]] \leq e^{-\epsilon^2 E[S]/2}, \tag{2}$$

where $E[S]$ is the mean of S. In our case, $m = s(n)$. $X_i = 1$ if the ith ball of the sample is red; otherwise $X_i = 0$. Note that $E(S) = s(n) \cdot L/n$ because we have $s(n)$ balls in the sample, and each ball is red with a probability equal to L/n.

It follows from inequality (1), letting $\epsilon = 1/2$, that

$$\textbf{Prob}[r \leq s(n)L/(2n)] \leq e^{-s(n)L/(12n)}.$$

Thus,

$$\textbf{Prob}[L/n \leq 2r/s(n)] \geq 1 - e^{-s(n)L/(12n)}.$$

It follows from inequality (2), letting $\epsilon = 1$, that

$$\textbf{Prob}[r > 2s(n)L/n] \leq e^{-s(n)L/(2n)}.$$

Thus,

$$\textbf{Prob}[L/n \geq r/(2s(n))] \geq 1 - e^{-s(n)L/(2n)}.$$

Thus, in our lemma, we should choose $c = 1/12$. □

In our case, red balls correspond to vertices in the separator, and we want to make sure that $L = O(n^{1-1/d})$. Thus as long as we sample more than $\Omega(n^{1/d} \log n)$ vertices, we can approximate the size of the separator to within a constant factor with very high probability (e.g., $1 - 1/n^2$). Once we can test the quality of a sphere, we can recursively build the initial fraction T_1 of an $O(n^{1-1/d})$-separator tree using the following procedure. Suppose $M = (A, xyz)$ is the given well-shaped mesh and Γ is its nearest neighborhood system. Let Γ' be a set of $n/\log n$ balls chosen uniformly at random from Γ, where n is the number of vertices of M.

ALGORITHM (initial tree(Γ')).

1. Let `found` = 0;
2. While (`found` == 0) do
   (a) Find a sphere separator S using the geometric separator algorithm.
   (b) Let $\Gamma_I = \Gamma_E = \Gamma_O = \emptyset$.
   (c) For each ball $B \in \Gamma'$
   - if B is in the vertex-separator of S then add B to Γ_O,
   - else if B is in the interior of S, then add B to Γ_I,
   - else add B to Γ_E.

   (d) We use Lemma 4.2 and $|\Gamma_O|$ to estimate the size of the vertex-separator induced by S. If the estimation shows that the size of the separator of S is small enough, then let `found` = 1.
3. Store S in the root of the tree.
4. If Γ_I has more than $O(\log^h n)$ balls, then call initial tree(Γ_I) to generate the left subtree.

5. If Γ_E has more than $O(\log^h n)$ balls, then call initial tree(Γ_E) to generate the right subtree.

Clearly, the above procedure takes expected $O(n)$ time. By Lemma 4.2, it finds a separator tree, denoted by T_1, which is the initial fraction of an $(n^{1-1/d})$-separator tree of the input mesh.

Remark 4.3. Another approach to construct an initial separator tree is to build the tree level by level. We choose independently a set of random balls for the construction of each level. This approach will takes $O(n)$ time if we choose no more than $O(n/\log^2 n)$ random balls for each level. Mathematically, the independence of each level makes it easier to apply Lemma 4.2. We will relate this approach with the multilevel partitioning scheme in section 6.

4.4. The identification problem if we know the separators. To complete the construction of the separator tree, we need to identify the submesh associated with each leaf of T_1. Notice that if we simply "push the input mesh down T_1" by comparing its elements with the sphere separators of T_1, then we will use $\Omega(n \log n)$ time; we need $O(\log n)$ time to identify each element or vertex.

The union of the vertex-separators used at the internal nodes of T_1 induces a multiway partition—a decomposition of the input mesh M into $|T_1|$-submeshes, one for each leaf of T_1; notice that there may be some leaves in T_1 whose submeshes are not connected. We will call this union the *multisector* associated with T_1. We now give an upper bound on the cardinality of the multisector associated with T_1.

LEMMA 4.4. *Let h be a positive integer. Let T_1' be the tree obtained from T_1 by removing all of its leaves. If T_1 has the property that the size of the submesh associated with each leaf of T_1' is $\Theta(\log^h n)$, then the cardinality of the multisector associated with T_1 is bounded from above by $O(n/\log^{(h/d)} n)$.*

Proof. By the assumption of the lemma, T_1 has at most $O(n/\log^h n)$ leaves. If the size of the submesh associated with an internal node of T_1 is m, then the size of the separator at the node is at most $O(m^{1-1/d})$ (by Theorem 3.1). Let $H(m)$ denote the maximum size of the multisector associated with a separator tree for a well-shaped mesh of m nodes. We have

$$H(m) = \begin{cases} H(\delta m) + H((1-\delta)m) + O(m^{1-1/d}) & \text{if } m > \log^h n, \\ 0 & \text{if } m \le \log^h n, \end{cases}$$

where $(d+1)/(d+2) \le \delta < 1$ is the constant given in Theorem 3.1. The solution of this recurrence, as shown in [32, 35], is $O(n/\log^{(h/d)} n)$. □

The following lemma shows that after removing the multisector associated with T_1, the input mesh is decomposed into at most $O(n/\log^{(h/d)} n)$ connected components.

LEMMA 4.5. *Let G be a connected graph of n vertices whose degree is bounded by Δ. Let s be any integer in the range $0 < s < n/\Delta$. Let S be any subset of s vertices of G. Then the number of connected components of G after the removal of S is bounded by $(\Delta - 1)s + 1$.*

Proof. We prove this lemma by an induction on s. Because G is connected and the degree G is bounded by Δ, the removal of any vertex of G can induce at most Δ connected components. Suppose the lemma is true for $s-1$ for $s > 1$. Without loss of generality, let $S = \{v_1, \ldots, v_s\}$. The removal of $\{v_1, \ldots, v_{s-1}\}$ results in at most $(\Delta-1)(s-1)+1$ connected components, say, $G_1, \ldots, G_t$, where $t \le (\Delta-1)(s-1)+1$. Suppose v_s is in G_i for an i in the range $1 \le i \le t$. The removal of v_s can induce at most Δ connected components from G_i. So the total number of components after removing S is at most $t + \Delta - 1 \le (\Delta-1)(s-1) + 1 + \Delta - 1 \le (\Delta-1)s + 1$. □

Because a well-shaped mesh has a constant degree and the cardinality of the multisector of T_1 is bounded by $O(n/\log^{(h/d)} n)$, after removing the multisector associated with T_1 the input mesh is decomposed into at most $O(n/\log^{(h/d)} n)$ connected components.

Suppose we have an "oracle" (or a "little magic bird") that always tells us the multisector T_1. Then we can apply a graph connectivity algorithm to identify all components of the mesh induced by the multisector associated with T_1. To identify the submesh associated with the leaves of T_1, we simply push one vertex from each component down the tree T_1 (by comparing its ball with T_1). If we choose $h > d$, then there are at most $O(n/\log n)$ components, and thus we only need $O(n)$ time.

Unfortunately, the problem of determining the multisector of a separator tree T_1 may be as hard as the problem of identifying the submeshes of the leaves of T_1!

4.5. Geometric simulation of graph connectivity. Recall from the previous subsection that the removal of the multisector of T_1 only induces $O(n/\log^{(h/d)} n)$ connected components. Our idea is to use this fact to reduce the number of balls to be pushed down T_1 in the process of identifying the submeshes of the leaves of T_1. For each component, we will push only one of its balls down T_1. We will then use a graph connectivity algorithm to determine the connected component that contains this vertex. To efficiently support this idea, we will call upon a geometric data structure due to Chazelle [7].

We first introduce some geometric notation. Suppose we have m hyperplanes $H = \{h_1, \ldots, h_m\}$ in $\mathbb{R}^d$. Note that H divides $\mathbb{R}^d$ into a collection of convex cells, called the *arrangement* of H. In general, there are $O(m^d)$ cells. For example, a set of $m/2$ horizontal lines and $m/2$ vertical lines in two dimensions divides $\mathbb{R}^2$ into $(m/2+1)^2$ rectangular cells. Let $\mathcal{A}(H)$ denote the arrangement of H (see [9] for the formal definition and properties of arrangements). In 1991, Chazelle[2] [7] constructed a data structure for $O(\log m)$ time point location in the arrangement of any set of m hyperplanes.

LEMMA 4.6 (see Chazelle [7]). *Let d be a positive constant integer. Let H be a set of m hyperplanes in $\mathbb{R}^d$. There is a data structure of size $O(m^d)$ that can be computed in $O(m^d)$ time to answer queries of the following type in $O(\log m)$ time: given a pair of points $p, q \in \mathbb{R}^d$, do p and q belong to the same cell in $\mathcal{A}(H)$?*

We will refer to the data structure in Lemma 4.6 as *Chazelle's structure* of $\mathcal{A}(H)$. We now show how to use Chazelle's structure in our setting.

Let w be a leaf of T_1. Let $w_1 = w$, w_2, w_3,...,$w_t = r$ be the path from w to the root r of T_1. Notice that $t = O(\log n)$. Let S_i, for an i in the range $2 \le i \le t$, be the separator sphere associated with w_i. Then a vertex v of the input mesh belongs to the submesh associated with w if for each i in the range $1 \le i \le t-1$, v is in the interior of S_{i+1} when w_i is the left child of w_{i+1} and v is in the exterior of S_{i+1} when w_i is the right child of w_{i+1}.

If we use a stereographic map to send the mesh onto the unit sphere in $\mathbb{R}^{d+1}$, then $S_2, \ldots, S_t$ are mapped to some hyperplanes $h_2, \ldots, h_t$ in $\mathbb{R}^{d+1}$ [30]. The interior and exterior of S_i are mapped to the two half spaces of h_i, respectively. Therefore, the stereographic image of the submesh associated with w is mapped to a cell of the arrangement of $h_2, \ldots, h_t$. We can use Chazelle's structure for $\mathcal{A}(h_1, \ldots, h_t)$ to support queries of the following type: does a given vertex v of the input mesh belong to the submesh associated with leaf w? The query takes $O(\log t) = O(\log\log n)$ time.

[2] A randomized version has been given by Clarkson.

Let $H_w = \{h_1, \ldots, h_t\}$. Let C_w be Chazelle's structure for $\mathcal{A}(H_w)$. Suppose we have already found a vertex v in the submesh associated with w. Then we can simulate a graph connectivity algorithm, a BFS-based algorithm of Tarjan [34], starting at v. The graph connectivity algorithm builds a breadth first search tree rooted at v by visiting v's neighbors and its neighbors' neighbors, etc. Suppose u is a neighbor of v in the original mesh. Then u belongs to the submesh associated with w if the stereographic images of u and v are contained in the same cell of $\mathcal{A}(H_w)$. We can test this condition in $O(\log\log n)$ time with the help of Chazelle's structure for $\mathcal{A}(H_w)$. If the connected component containing v has L vertices, we can determine its vertices in $O(L\log\log n)$ time.

Therefore, for each leaf w of T_1, we build its Chazelle's structure C_w. We use $O(|T_1|\log^{d+1} n)$ time for all leaves of T_1. If we choose $h \geq d+1$, then $|T_1| \leq n/\log^h n \leq n/\log^{d+1} n$. Hence, we use not more than linear time.

We now give our algorithm for determining the submeshes associated with the leaves of T_1.

ALGORITHM (submesh identification).

Input A well-shaped mesh $M = (A, xyz)$ in $\mathbb{R}^d$ and a tree of spheres T_1.

1. Build Chazelle's structure C_w for each leaf w of T_1.
2. Label all nodes in M "unidentified."
3. If there is an unidentified node v, then push the ball of v down T_1. If v is identified to be a member of the multisector associated with T_1, label it "multisector." Return to the beginning of this step. Otherwise, v reaches some leaf w of T_1 and we label v "identified with w."
4. Simulate a graph connectivity algorithm (a BFS-based algorithm [34]) starting from v to identify the component that contains v. A neighbor of v in the mesh is not in the same component of v iff its stereographic image does not belong to the same cell in $\mathcal{A}(H_w)$. So the basic step of the connectivity algorithm can be performed in $O(\log\log n)$ time with the help of Chazelle's structure for $\mathcal{A}(H_w)$.
5. After identifying the component that contains v, return to the beginning of Step 3.

It follows from Lemma 4.4 that if $h > d+1$ then the procedure above takes $O(n\log\log n)$ time. Therefore, we have proved Theorem 4.1.

5. Applications. The separator tree decomposition of a graph has many applications, especially for the design of efficient divide-and-conquer algorithms. In this section, we present two applications: (1) sparse matrix ordering and symbolic factorization and (2) point location in a 3D well-shaped mesh. The latter is useful in hierarchical methods [36].

5.1. Nested dissection. Nested dissection was first proposed by George [12] for solving linear systems on 2D regular grids. Lipton, Rose, and Tarjan [26] extended the method to general planar systems using a result of Lipton and Tarjan [27] that every planar graph has $\sqrt{8n}$ separators. The separator result of Miller et al. (Theorem 3.1) can be used to extend the nested-dissection method to 3D linear systems [17]. As shown in [31, 35], $O(n^{1-1/d})$ is the best possible separator bound for d-dimensional meshes. As a consequence of a result of Agrawal and Klein [1], the application of Theorem 3.1 to nested-dissection generates a provably good elimination ordering for

3D linear systems.

A nested-dissection ordering of a symmetric sparse matrix A can be found in linear time given a separator tree for $G(A)$. We simply traverse the separator tree in postorder to generate the ordering. Thus, for sparse matrices whose graphs are well-shaped meshes, our algorithm gives a randomized $O(n \log\log n)$ time construction of a provably good nested-dissection ordering. The amount of fill is bounded by $O(n^{4/3})$ and the elimination tree height is $O(n^{2/3})$ for 3D meshes.

The symbolic factorization step is to predict the pattern of fill in the process of numerical elimination. In doing so, it allocates the necessary amount of space to accommodate fill in the numerical elimination. Gilbert, Ng, and Peyton [15] showed that from a separator tree the *elimination tree* can be found in $O(n\alpha(n))$ time, where $\alpha(n)$ is the inverse of the Ackerman function. They also showed that with the elimination tree one can count fill (and get all the vertex degrees in the filled graph) in another $O(n\alpha(n))$ time.

For proofs of these statements and many other fascinating facts about elimination trees and filled graphs consult [14, 15, 16, 28, 29].

Therefore, we have the following corollary.

COROLLARY 5.1. *If the graph of a symmetric sparse matrix A is a well-shaped mesh, then a provably good nested-dissection ordering of A, its elimination tree, and its fill information can be found in $O(n \log\log n)$ time.*

5.2. 3D point location. One of the key problems in hierarchical methods is the problem of point location in a well-shaped mesh; that is, given a point p we would like to quickly determine which element of the mesh contains it [6, 36]. Point location is used for interpolation between two neighboring meshes in the multilevel discretization.

The geometric mesh partitioner generates a sphere S that divides the elements F of the input mesh (A, xyz) into three subsets F_I, F_E, F_O such that $\max(|F_I|, |F_E|) \leq \delta|F|$ for some constant δ and $|F_O| = O(|F|^{1-1/d})$.

We observe that S can be used to prune the region for point location. If a point p is in the interior of S, then there is no need to search p in F_E; hence a single comparison with S eliminates a constant fraction of the region to be searched. Similarly, if p is in the exterior of S, then there is no need to search p in F_I.

We can pursue this idea recursively to design a data structure. We build a binary tree (as in [32, 35]). The tree is constructed recursively. The root of the tree contains the information of S, the top-level separating sphere. The root has two subtrees, one recursively generated for $F_I \cup F_O$ and the other generated for $F_E \cup F_O$. The recursive construction stops when the number of elements is less than some constant. Because each internal node uses a balanced sphere separator, the tree has $O(\log n)$ levels. Also, because $|F_O|$ is sublinear in $|F|$, as shown in [35], the tree uses only $O(n)$ space. To locate a point p, we push the point down the tree. The query time is clearly proportional to the height of the tree, which is $O(\log n)$. The same point location structure works for any fixed dimension as long as the underlying mesh is well shaped [32, 35].

If we sample the elements of the mesh in the step of testing quality, the same construction of section 4 can be shown to have random $O(n \log\log n)$ time complexity. This improves the previous result by a factor of $O(\log n/\log\log n)$.

COROLLARY 5.2. *Given a well-shaped mesh $M = (A, xyz)$ in $\mathbb{R}^d$, we can construct an $O(\log n)$ query time and linear space point location structure for M in $O(n \log\log n)$ time.*

6. Final remarks and open questions. The results of this paper are mathematical and algorithmic in nature. However, some techniques developed here may have potential applications to practical implementations.

In this paper, we have shown that geometric sampling can be used to speed up nested-dissection and graph partitioning algorithms. At a high level, a good sampling technique allows us to solve the original problem by first solving a smaller problem and then projecting the solution of the smaller problem back to the original problem. We have proved that random sampling can be used to preserve the quality of the separator decomposition for well-shaped meshes. In practice, we can combine the techniques of this paper with some other commonly used quality enhancement heuristics.

One-level geometric sampling-based scheme.

1. Use the geometric sample of the input mesh to construct an initial multiway partition of the mesh.
2. Project the multiway partition back to the original mesh to obtain an initial nested-dissection ordering; the projection can be performed by nested dissection (as presented in this paper) or minimum degree on the submeshes associated with the leaves of the initial separator tree.
3. Apply iterative techniques such as Kernighan–Lin [24] to enhance the quality of the partition and ordering.

We can recursively apply our sampling technique to obtain a geometric sampling-based multilevel scheme for nested-dissection ordering and multiway partitioning.

Geometric sampling-based multilevel scheme.

Input A well-shaped mesh $M = (A, xyz)$ in $\mathbb{R}^d$ and its nearest neighborhood system Γ.

1. Choose a sequence of random sample $\Gamma_1, \ldots, \Gamma_t$ from Γ such that (1) $|\Gamma_1| = \Theta(n^{1/d} \log n)$, (2) $|\Gamma_i| = 2|\Gamma_{i-1}|$ for each i in the range $2 \leq i \leq t$, and (3) $|\Gamma_t| = n/2$. Note that $t = O(\log n)$.
2. Construct the initial separator tree and multiway partition for Γ_1.
3. For $i = 1$ to $t-1$, project the separator tree from Γ_i to Γ_{i+1} and apply Kernighan–Lin or other heuristics to enhance the quality of the partition.
4. Project the separator tree from Γ_t to M and apply Kernighan–Lin or other heuristics to enhance the quality of the partition and ordering.

The choice of $\Theta(n^{1/d} \log n)$ is justified by Lemma 4.2. See the paragraph right after the proof of Lemma 4.2. Associated with each Γ_i there is a graph, called the *overlap graph*, which is implicitly defined in Theorem 2.2. Let α be the constant given in Theorem 2.2. The α overlap graph G_i of Γ_i defines an edge between two balls B_u and B_v in Γ_i if B_u intersects $\alpha \cdot B_v$ and $\alpha \cdot B_u$ intersects B_v. We can view G_i as a coarsened graph of M as well as a coarsened graph of G_{i+1}. In other words, we use geometric sampling to perform graph coarsening in the multilevel partitioning and ordering schemes.

Our proofs in this paper (beginning with Lemma 4.2) can be easily extended to show that our multilevel scheme always generates an $O(n^{1-1/d})$-separator tree decomposition and hence gives a provably good nested-dissection ordering and multiway partition for well-shaped meshes in d dimensions. See Remark 4.3.

COROLLARY 6.1. *Given a well-shaped mesh $M = (A, xyz)$ in $\mathbb{R}^d$, with high probability, the geometric sampling-based multilevel scheme finds an $O(n^{1-1/d})$-separator*

tree decomposition of M.

It is an interesting project to implement and compare this geometric sampling-based multilevel scheme with some other multilevel schemes that have been proposed and implemented by Barnard and Simon [3], Hendrickson and Leland [21], and Karypis and Kumar [23]. These three multilevel schemes (especially the one by Karypis and Kumar [23]) have produced very good experimental results, although it is still open whether mathematically they always provide provably good partitions and orderings. It is also an interesting research direction to understand the connection between graph sampling and graph coarsening. Graph coarsening is a key problem for multilevel computations.

Our improved algorithm can be parallelized optimally. Our construction, in conjunction with the parallel algorithm of Frieze, Miller, and Teng [11], yields a parallel algorithm that runs in parallel $O(\log n)$ time on $n \log\log n / \log n$ processors.

Recently, Klein et al. [25] presented a linear time algorithm for the single-source shortest path problem for planar graphs. They used the separator tree decomposition algorithm of Goodrich [19]. Our result in conjunction with theirs [25] yields a randomized $O(n \log\log n)$ time algorithm for a large class of geometric graphs such as well-shaped meshes and nearest neighbor graphs in any fixed dimension.

COROLLARY 6.2. *The single-source shortest path problem for well-shaped meshes and nearest neighbor graphs can be solved in random $O(n \log\log n)$ time.*

We conclude this paper with some algorithmic open questions.

1. Can we construct an $O(n^{1-1/d})$-separator tree in linear time?
2. Can we construct an $O(n^{1-1/d})$-separator tree in deterministic $o(n \log n)$ time?

Acknowledgments. I would like to thank Steve Vavasis for motivating me to work on this problem and Ken Clarkson for the discussion of the geometric search structure of Chazelle. I would like to thank John Gilbert, Gary Miller, and Dan Spielman for helpful discussions. I would like to thank Cleve Ashcraft for carefully reading an earlier version of this paper and for his technical and editorial comments and suggestions that greatly improved this paper.

REFERENCES

[1] A. AGRAWAL AND P. KLEIN, *Cutting down on fill using nested dissection: Provably good elimination orderings*, in Sparse Matrix Computations: Graph Theory Issues and Algorithms, A. George, J. Gilbert, and J. Liu, eds., IMA Volumes in Mathematics and its Applications 56, Springer-Verlag, New York, 1993, pp. 31–56.

[2] N. ALON, J. H. SPENCER, AND P. ERDÖS, *The Probabilistic Method*, John Wiley, New York, 1992.

[3] S. T. BARNARD AND H. D. SIMON, *A fast multilevel implementation of recursive spectral bisection for partitioning unstructured problems*, Concurrency: Practice and Experience, 6 (1994), pp. 101–117.

[4] M. BERN AND D. EPPSTEIN, *Mesh generation and optimal triangulation*, in Computing in Euclidean Geometry, D.-Z. Du and F. Hwang, eds., Lecture Notes Series on Computing, Vol. 1., World Scientific, River Edge, NJ, 1992, pp. 23–89.

[5] M. BERN, D. EPPSTEIN, AND J. R. GILBERT, *Provably good mesh generation*, in 31st Annual Symposium on Foundations of Computer Science, IEEE, St. Louis, MO, 1990, pp. 231–241.

[6] T. F. CHAN AND B. SMITH, *Domain decomposition and multigrid algorithms for elliptic problems on unstructured meshes*, Contemp. Math., 180 (1993), pp. 1–14.

[7] B. CHAZELLE, *Cutting hyerplanes for divide and conquer*, Discrete Comput. Geom., 9 (1991), pp. 145–158.

[8] K. CLARKSON, D. EPPSTEIN, G. L. MILLER, C. STURTIVANT, AND S.-H. TENG, *Approximating center points with iterated Radon points*, in Proc. of 9th ACM Symposium on Computational Geometry, San Diego, CA, 1993, pp. 91–98.

[9] H. Edelsbrunner, *Algorithms in Combinatorial Geometry*, Springer-Verlag, New York, 1987.
[10] I. Fried, *Condition of finite element matrices generated from nonuniform meshes*, AIAA J., 10 (1972), pp. 219–221.
[11] A. M. Frieze, G. L. Miller, and S.-H. Teng, *Separator based divide and conquer in computational geometry*, in Proc. of the 1992 ACM Symposium on Parallel Algorithms and Architectures, San Diego, CA, 1992, pp. 420–430.
[12] J. A. George, *Nested dissection of a regular finite element mesh*, SIAM J. Numer. Anal., 10 (1973), pp. 345–363.
[13] A. George and J. W. H. Liu, Computer Solution of Large Sparse Positive Definite Systems, Prentice–Hall, Englewood Cliffs, NJ, 1981.
[14] J. R. Gilbert, *Predicting structure in sparse matrix computations*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 62–79.
[15] J. R. Gilbert, E. G. Ng, and B. W. Peyton. *An efficient algorithm to compute row and column counts for sparse Cholesky factorization*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1075–1091.
[16] J. R. Gilbert and R. E. Tarjan, *The analysis of a nested dissection algorithm*, Numer. Math., 50 (1987), pp. 377–404.
[17] J. R. Gilbert, G. L. Miller, and S.-H. Teng, *Geometric mesh partitioning and nested dissection*, in 12th Householder Symposium on Numerical Algebra, Plenary talk, Los Angeles, CA, 1993.
[18] J. R. Gilbert, G. L. Miller, and S.-H. Teng, *Geometric mesh partitioning: Implementation and experiments*, SIAM J. Sci. Comput., submitted.
[19] M. Goodrich, *Planar separators and parallel polygon triangulation*, in Proc. of the 24th Annual ACM Symposium on the Theory of Computing, Montreal, Quebec, Canada, 1994, pp. 507–516.
[20] T. Hagerup and C. Rüb, *A guided tour of Chernoff bounds*, Inform. Process. Lett., 33 (1990), pp. 305–308.
[21] B. Hendrickson and R. Leland, *A Multilevel Algorithm for Partitioning Graphs,* Tech. report 93-1301, Sandia National Labs, Albuquerque, NM, 1993.
[22] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, Amer. Statist. Assoc. J., 58 (1963), pp. 13–29.
[23] G. Karypis and V. Kumar, *A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs*, Tech. report 95-035, Department of Computer Science, University of Minnesota, Minneapolis, MN, 1995.
[24] B. W. Kernighan and S. Lin, *An efficient heuristic procedure for partitioning graphs*, Bell Sys. Tech. J., 49 (1970), pp. 291–307.
[25] P. Klein, S. Rao, M. Rauch, and S. Subramanian *Faster shortest-path algorithms for planar graphs*, in Proc. of the 26th Annual ACM Symposium on the Theory of Computing, Montreal, Quebec, Canada, 1994, pp. 27–37.
[26] R. J. Lipton, D. J. Rose, and R. E. Tarjan, *Generalized nested dissection*, SIAM J. Numer. Anal., 16 (1979), pp. 346–358.
[27] R. J. Lipton and R. E. Tarjan, *A separator theorem for planar graphs*, SIAM J. Appl. Math., 36 (1979), pp. 177–189.
[28] J. W. H. Liu, *A compact row storage scheme for cholesky factors using elimination trees*, ACM Trans. Math. Software, 12 (1986), pp. 127–148.
[29] J. W. H. Liu, *The role of elimination trees in sparse factorization*, SIAM J. Matrix Anal., 11 (1990), pp. 134–173.
[30] G. L. Miller, S.-H. Teng, W. Thurston, and S. A. Vavasis, *Automatic mesh partitioning*, in Sparse Matrix Computations: Graph Theory Issues and Algorithms, A. George, J. Gilbert, and J. Liu, eds., IMA Volumes in Mathematics and its Applications 56, Springer-Verlag, New York, 1993, pp. 57–84.
[31] G. L. Miller, S.-H. Teng, W. Thurston, and S. A. Vavasis, *Geometric separators for finite element meshes*, SIAM J. Sci. Comput., 1997, to appear.
[32] G. L. Miller, S.-H. Teng, W. Thurston, and S. A. Vavasis, *Separators for sphere-packings and nearest neighborhood graphs*, J. ACM, 44 (1997), pp. 1–29.
[33] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Prentice–Hall, Englewood Cliffs, NJ, 1973.
[34] R. E. Tarjan, *Data Structures and Network Algorithms*, SIAM, Philadelphia, PA, 1985.
[35] S.-H. Teng, *Points, Spheres, and Separators: A Unified Geometric Approach to Graph Partitioning*, Ph.D. thesis, Carnegie-Mellon University, School of Computer Science, Pittsburgh, PA, 1991.
[36] S.-H. Teng. *Geometric approaches to hierarchical and adaptive computing*, in 5th SIAM Conference on Applied Linear Algebra, SIAM, Philadelphia, PA, 1994, pp. 51–57.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 3, pp. 566–589, July 1997

004

AN EFFICIENT IMPLEMENTATION OF THE NONSYMMETRIC LANCZOS ALGORITHM*

DAVID DAY†

Abstract. Lanczos vectors computed in finite precision arithmetic by the three-term recurrence tend to lose their mutual biorthogonality. One either accepts this loss and takes more steps or re-biorthogonalizes the Lanczos vectors at each step. For the symmetric case, there is a compromise approach. This compromise, known as maintaining semiorthogonality, minimizes the cost of reorthogonalization. This paper extends the compromise to the two-sided Lanczos algorithm and justifies the new algorithm.

The compromise is called *maintaining semiduality*. An advantage of maintaining semiduality is that the computed tridiagonal is a perturbation of a matrix that is exactly similar to the appropriate projection of the given matrix onto the computed subspaces. Another benefit is that the simple two-sided Gram–Schmidt procedure is a viable way to correct for loss of duality.

A numerical experiment is included in which our Lanczos code is significantly more efficient than Arnoldi's method.

Key words. Lanczos algorithm, breakdown, sparse eigenvalue problems, biorthogonalization methods

AMS subject classification. 65F15

PII. S0895479895292503

1. Introduction.

For non-Hermitian matrices approximate eigenvalues from the (two-sided) Lanczos process are much more accurate (for the same elapsed time and starting vectors) than those from the Arnoldi method. Consequently, it is important to implement the Lanczos algorithm as well as possible. This paper summarizes the analysis in [7] and claims to show the best (or nearly best) way to do it.

This article shows that it is not necessary to re-biorthogonalize the Lanczos vectors at every step to approximate the behavior of the algorithm in exact arithmetic. A property of the computed Lanczos vectors called *semiduality* may be imposed (defined in section 3) and suffices for keeping close to the exact algorithm at minimal cost. Semiduality is less expensive to maintain than duality, yet equally effective. Having stated our contribution, we resume the introduction.

Krylov subspace methods determine a useful basis for the Krylov subspace

$$\mathcal{K}_i(q, B) = \text{span}(q, Bq, \ldots, B^{i-1}q).$$

The eigenvalues of certain projections of B onto $\mathcal{K}_i(q, B)$ serve as approximations to eigenvalues of B. The eigenvalues of projections of B are often called Ritz values. They are not Raleigh–Ritz approximations, except in the Hermitian case, but we do not have a better name for them.

*Received by the editors September 27, 1995; accepted for publication (in revised form) by M. Gutknecht July 9, 1996.

http://www.siam.org/journals/simax/18-3/29250.html

†Applied and Numerical Math, Sandia National Laboratories, Albuquerque, NM 87185-5800 (dday@cs.sandia.gov). The author gratefully acknowledges support from ONR contract N00014-90-J-1372 and the Sandia National Laboratory AMS Fellowship.

The most popular Krylov subspace method for non-Hermitian matrices is the Arnoldi algorithm [27]. An orthonormal basis of $\mathcal{K}_i(q, B)$ is computed. The orthogonal projection of B onto $\mathcal{K}_i(q, B)$ is represented by an $i \times i$ Hessenberg matrix.

The non-Hermitian or two-sided Lanczos algorithm is another Krylov subspace method. Given two starting vectors $p^* = p_1^*$ and $q = q_1$, the two-sided Lanczos algorithm simultaneously computes a basis for the right Krylov subspace $\mathcal{K}_i(q, B)$ and a dual basis for the left Krylov subspace

$$\mathcal{K}^i(p^*, B) = \text{span}(p^*, \ldots, p^* B^{i-1}).$$

The Lanczos algorithm computes the partial reduction of B to tridiagonal form.

In exact arithmetic the Hermitian Lanczos algorithm determines a matrix of orthogonal vectors Q, while the two-sided Lanczos algorithm determines two matrices P and Q such that P^*Q is diagonal. This relation among the Lanczos vectors is often called *biorthogonality* [16].

Lanczos vectors computed in finite precision arithmetic by the three-term recurrence tend to lose their mutual biorthogonality. Two ways to compensate for this phenomenon are known: Lanczos with full re-biorthogonalization (LanFRB) [4] and an acceptance of the loss of biorthogonality which forces more steps to be taken [6, 10]. For the Hermitian case, a compromise is known [11, 19, 20, 28, 29]. This compromise, known as maintaining semiorthogonality, minimizes the number of reorthogonalizations. This article extends the compromise to the two-sided Lanczos algorithm and justifies the new algorithm.

There are known cases when the simple recurrence takes extreme amounts of time [10]. On the other hand, LanFRB is expensive for a long run. The compromise we present in this article is better than either of the two extremes.

Better approximations to eigenvalues of B tend to be computed from a single Krylov subspace of dimension 100 than from four Krylov subspaces of dimension 25. To take advantage of this property, we want to use large Krylov subspaces. The amount of data transfer (from memory to the computational unit) required in Lanczos with full re-biorthogonalization when n is large is significant. In this respect, the compromise is at least twice as fast as maintaining full biorthogonality. Usually it is much faster. See section 6.2.

The state-of-the-art in Lanczos methods for eigenvalue problems is to select from one of four algorithms. For linear solvers there are many more options: two-term versus three-term, CGS, BiCGStab1/2/e, QMR, TFQRM, and so on. But for eigenvalue problems the user first selects either the three-term recurrence or LanFRB. This choice is a trade-off between the low cost per step of the three-term recurrence and the limited number of Lanczos steps taken by LanFRB. Then the user selects an implementation with or without look-ahead [10, 22, 21]. Look-ahead enhances stability while increasing cost modestly. This article does not consider implementations with the look-ahead feature.

The word reorthogonalization in the Hermitian case is ugly enough, but the analogous term re-biorthogonalization goes too far (nine syllables). So we seek a term with fewer syllables. In functional analysis row vectors represent linear functionals and the property $p_i^* q_j = \delta_{ij}$ (Kronecker's delta) says that the ordered sets $\{p_1^*, \ldots, p_j^*\}$ and $\{q_1, \ldots, q_j\}$ are a pair of dual bases for $\mathcal{K}_i(q, B)$ and $\mathcal{K}^i(p^*, B)$. So we use the term dual instead of biorthogonal. Consequently, we speak of maintaining duality, local duality, and semiduality (introduced in section 3).

1.1. Summary. Our results extend earlier work done in the Hermitian case [28, 29], but new issues arise in the non-Hermitian case. In exact arithmetic the Lanczos algorithm determines a tridiagonal–diagonal pencil (T, Ω) such that $\Omega^{-1}T$ is similar to the projection of B onto the spans of the Krylov subspaces. See Definition 2.1. The diagonal elements of Ω are defined to be the inner products of consecutive pairs of normalized Lanczos vectors. In exact arithmetic the algorithm breaks down if $\Omega = \text{diag}(\omega_i)$ is singular. In finite precision arithmetic breakdowns are rare, but near breakdowns are not. It is tempting to require that $|\omega_i| \geq \sqrt{\epsilon}$, where ϵ is the round-off unit, but the rather lengthy analysis of [7] shows that the algorithm is still viable provided that $|\omega_j| \geq (n+10j)\epsilon$. Below that level the accuracy of the Ritz values does not generally improve if the recurrence continues.

The remainder of this work is organized as follows. Section 2 contains a discussion of what is known about solving eigenvalue problems using the two-sided Lanczos process. The basic properties of the Lanczos algorithm are reviewed, the implementation of the three-term recurrence is outlined, convergence theory is discussed, and our practical experience with implementations of the three-term recurrence is summarized.

With that done, we move on to the tricky issue of when to "re-biorthogonalize" or correct the Lanczos vectors to restore duality. The candidate Lanczos vectors are computed by the three-term recurrence, but at certain steps the loss of duality of the candidate Lanczos vectors to the previous Lanczos vectors is "too large," and then we correct them to obtain the final Lanczos vectors. To obtain a competitive algorithm, correction steps are implemented just like a step of LanFRB. Since the duality of the Lanczos vectors is not maintained to full precision, this process must be justified. The viability of the two sided Gram–Schmidt process is established in section 3.

In section 4 the properties of the Lanczos algorithm *with correction* are developed. In section 4.4 we show how to monitor the loss of duality among the computed Lanczos vectors without significantly increasing the cost of the algorithm. For efficiency the correction steps must be invoked as rarely as possible consistent with maintaining accuracy in the approximations.

In section 5 we prove that an added advantage of maintaining semiduality is that the computed pencil (T, Ω) is a perturbation of a pencil that is exactly equivalent to the projection of the operator onto the computed subspaces. The norm of the perturbation is as small as the data warrants. Section 6 illustrates some of our results with some challenging numerical examples.

2. Two-sided Lanczos. The two-sided Lanczos algorithm is based on the partial reduction of a non-Hermitian matrix B to tridiagonal form. The Lanczos algorithm starts from an arbitrary pair of vectors $p^* = p_1^*$ and $q = q_1$. After j successful steps, the matrices P_j^* and Q_j are produced. The rows of P_j^* span the Krylov subspace $\mathcal{K}^j(p^*, B)$ and the columns of Q_j span $\mathcal{K}_j(q, B)$. The matrix $T_j = P_j^* B Q_j$ is tridiagonal; $\Omega_j = P_j^* Q_j$ is diagonal. In finite precision arithmetic the latter will no longer be true, and we will set $\Omega_j = \text{diag}(P_j^* Q_j)$.

Certain implementations scale the Lanczos vectors so that $\Omega = I$ [1, 37], and others maintain the unit length of all the Lanczos vectors [4, 10, 22]. Our analysis of the Lanczos algorithm requires that the unit length of all the Lanczos vectors be maintained. Normalizing the Lanczos vectors is necessary in this work because for the resulting more complicated algorithm it is possible to establish certain properties of the quantities computed in finite precision arithmetic which are required to justify the Lanczos algorithm with correction (see [7]); this would be impossible based on less

precise models such as [1]. A useful result of our work is that Ω can become nearly singular, $\text{cond}(\Omega) = \mathcal{O}(1/\epsilon)$, without spoiling the algorithm.

The eigenvalues of an *oblique projection* of B are used to approximate the eigenvalues of B.

DEFINITION 2.1. *Let $Q_j = [q_1, \ldots, q_j]$ and $P_j^* = [p_1, \ldots, p_j]^*$ have full rank. If $P_j^* Q_j$ is invertible then*

$$\Pi_j = Q_j \Omega_j^{-1} P_j^*$$

is a projector ($\Pi_j^2 = \Pi_j$). It is not orthogonal ($\Pi_j^* \neq \Pi_j$) in general. We say that Π_j is an oblique projector onto $\text{Range}(Q_j)$. *It is also an oblique projector onto the dual space $\{u^* \Pi_j : u \in \mathbf{C}^j\} = \text{Range}(P_j)^*$. Thus $\Pi_j B \Pi_j$ is a projection of B onto the pair* $\text{Range}(Q_j)$ *and* $\text{Range}(P_j)^*$.

Assuming that Q_n and P_n^* exist (that is, the algorithm does not break down), the representation of B with respect to the basis $\{q_1, \ldots, q_n\}$ is $Q_n^{-1} B Q_n$ and we have $\Pi_n = I$, which implies that $Q_n^{-1} = \Omega_n^{-1} P_n^*$ and $Q_n^{-1} B Q_n = \Omega_n^{-1} T_n$. The tridiagonal $\Omega_j^{-1} T_j$ represents $\Pi_j B \Pi_j$ in the dual bases $\{q_1, \ldots, q_j\}$ and $\{\omega_1^{-1} p_1^*, \ldots, \omega_j^{-1} p_j^*\}$. Similarly, the representation corresponding to P_j^* and $Q_j \Omega_j^{-1}$ is $T_j \Omega_j^{-1}$.

2.1. The three-term recurrences. The Lanczos vectors satisfy a pair of three-term recurrences

$$\beta_{i+1} p_{i+1}^* = p_i^* B - \frac{\alpha_i}{\omega_i} p_i^* - \frac{\gamma_i \omega_i}{\omega_{i-1}} p_{i-1}^* \tag{2.1}$$

and

$$q_{i+1} \gamma_{i+1} = B q_i - q_i \frac{\alpha_i}{\omega_i} - q_{i-1} \frac{\beta_i \omega_i}{\omega_{i-1}}. \tag{2.2}$$

The coefficients α_i and ω_i are chosen so that the right-hand side of (2.1) annihilates $q_1, \ldots, q_i$ and the right-hand side of (2.2) is annihilated by $p_1^*, \ldots, p_i^*$. The βs and γs come from the normalizing convention. The recurrence stops if $\beta_{j+1} \omega_{j+1} \gamma_{j+1} = 0$. With

$$T_j := \text{tridiag} \begin{pmatrix} & \beta_2 \omega_2, & \cdots & , \beta_j \omega_j \\ \alpha_1, & & \cdots & & , \alpha_j \\ & \gamma_2 \omega_2, & \cdots & , \gamma_j \omega_j & \end{pmatrix},$$

equations (2.1) and (2.2) may be written in compact form:

$$P_j^* B - T_j \Omega_j^{-1} P_j^* = e_j \beta_{j+1} p_{j+1}^* \tag{2.3}$$

and

$$B Q_j - Q_j \Omega_j^{-1} T_j = q_{j+1} \gamma_{j+1} e_j^*, \tag{2.4}$$

where $e_j = (0, \ldots, 0, 1)^*$.

2.2. Ritz triplet convergence. The eigenvalues of B are approximated using the eigenvalues of the pair (T_j, Ω_j) for increasing j. Given an eigentriplet $(u_i^{(j)*}, \theta_i^{(j)}, v_i^{(j)})$,

$$u_i^{(j)*} T_j = \theta_i^{(j)} u_i^{(j)*} \Omega_j \quad \text{and} \quad T_j v_i^{(j)} = \Omega_j v_i^{(j)} \theta_i^{(j)},$$

form the Ritz triplet $(x_i^{(j)*}, \theta_i^{(j)}, y_i^{(j)})$ where

$$x_i^{(j)*} = u_i^{(j)*} P_j^* \quad \text{and} \quad y_i^{(j)} = Q_j v_i^{(j)}. \tag{2.5}$$

Ritz triplets approximate eigentriplets of B. In discussions of the analysis of the quantities computed after j Lanczos steps, we omit the superscript (j) for clarity.

The expression $x_i^* \times (2.4) \times v_i$ reduces to

$$x_i^* B y_i = \theta_i x_i^* y_i.$$

In other words, θ_i is the *generalized* Rayleigh quotient corresponding to $x_i^* = u_i^* P_j^*$ and $y_i = Q_j v_i$. See section 11 of [18] for a discussion of generalized Rayleigh quotients.

We now list what is known about the approximations derived from the first j steps of the algorithm.

Multiply (2.4) by v_i from the right and substitute (2.5) to obtain

$$B y_i - y_i \theta_i = q_{j+1} \gamma_{j+1} v_i(j), \tag{2.6}$$

where $v_i(j)$ is the jth component of $v_i = v_i^{(j)}$. The remarkable property of (2.6) is that the right-hand side can be computed without forming y_i. Even in exact arithmetic $\|y_i\|_2$ can be smaller than $\|v_i\|_2$. Thus a small value of $|v_i(j)|$ is a necessary though not sufficient indication that θ_i is close to an eigenvalue of B.

The perturbation theory for the eigenvalue problem is more complicated than in the Hermitian case. The Lanczos algorithm eventually yields approximate eigentriples $(\hat{x}_i^*, \theta_i, \hat{y}_i)$, where $\|\hat{x}_i^*\|_2 = 1 = \|\hat{y}_i\|_2$ such that the corresponding residuals $\|\hat{x}^*(B - \theta I)\|_2$ and $\|(B - \theta I)\hat{y}\|_2$ are small. Such triples exactly solve a nearby eigenvalue problem [14]. The good thing is that the eigenvalues of $\Omega_j^{-1} T_j$ for which the residual norms are small persist as approximate eigenvalues of $\Omega_k^{-1} T_k$ for $k > j$ [14].

The residual norm $\|(B - \theta_i I)\hat{y}_i\|_2$ is a pessimistic estimate of the accuracy of θ_i and a good estimate of the accuracy of $\hat{y}_i$. The accuracy of generalized Rayleigh quotients is proportional to the product of the residual norms. To be precise, if $(\hat{x}_i^*, \theta_i, \hat{y}_i)$ approximates an eigentriple of B that is well separated (see Theorem 2.1 in section 5 of [31]), then the accuracy of θ_i is proportional to

$$\|\hat{x}_i^*(B - \theta_i I)\|_2 \|(B - \theta_i I)\hat{y}_i\|_2.$$

This product divided by

$$\operatorname{gap}(\theta_i, T_j) = \min_{k \neq i} |\theta_i - \theta_k|$$

appears to be a realistic backward error estimate for θ_i [3].

One-sided algorithms, and in particular the Arnoldi algorithm, do not enjoy this property.

To factor the exact shrinkage $\|x\|_2/\|u\|_2$ and $\|y\|_2/\|v\|_2$ into the error estimates to obtain asymptotic error bounds, one must first compute the Ritz triplets. For an $n \times n$ real operator B, after j Lanczos steps the number of real floating point operations (flops) required to compute the matrices of left and right eigenvectors for m Ritz values is $8nmj$. This is often more flops than are required for the Lanczos run.

Fortunately, a realistic lower bound on the shrinkage is available if the duality of the computed Lanczos vectors is maintained. In this case $y = Q_j v$ satisfies $P_j^* y =$

$\Omega_j v$. Also, since the Lanczos vectors are normalized to have unit Euclidean length, $\|P_j^*\|_2 \le \sqrt{j}$. Combine these two equations to find

$$\|y\|_2 \ge \frac{\|P_j^*\|_2}{\sqrt{j}}\|y\|_2 \ge \frac{\|P_j^* y\|_2}{\sqrt{j}} = \frac{\|\Omega_j v\|_2}{\sqrt{j}}.$$

Similarly, if $x^* = u^* P_j^*$, then $\|x^*\|_2 \ge \|u^*\Omega_j\|_2/\sqrt{j}$.

2.3. Practical experience. Without careful observation, there is no science. This section discusses surprising behavior that has been consistently observed in large-scale scientific computations using the non-Hermitian Lanczos algorithm [10]. In the case $p_1 = q_1$,

- the sequence $\{|\omega_i|\}_{i>0}$ tends to be decreasing, sometimes precipitously,
- even when $|\omega_j|$ has declined to nearly $10n\epsilon$, approximate eigenvalues, eigenvectors, and solutions to linear systems continue to converge,
- even when $|\omega_j|$ has declined to nearly $10n\epsilon$, (T_j, Ω_j) is almost always graded so that the growth factor

$$\Phi_j = \max(\|T_j\Omega_j^{-1}\|_\infty, \|\Omega_j^{-1}T_j\|_1)/\|B\|_2 \tag{2.7}$$

  is of order unity, say less than 10.

As usual, ϵ denotes the machine precision. The scalar Φ_j measures the relative size of the intermediate quantities introduced during the Lanczos algorithm. It is essential to distinguish

$$\|\Omega_j^{-1}\|_2 = 1/\min_i |\omega_i|$$

from Φ_j. The quantity $\|\Omega_j^{-1}\|_2$ can be large, nearly ϵ^{-1}, without necessarily affecting the accuracy of the eigenvalues and eigenvectors. The error in computing the eigenvalues of $\Omega_j^{-1}T_j$ is, among other things, proportional to $\|\Omega_j^{-1}T_j\|_1$. But in the rare case that Φ_j is large, the Lanczos algorithm is unstable due to the introduction of large intermediate quantities. The approximate eigenvalues suffer perturbations like $\epsilon\Phi_j\|B\|_2$. In this case, look-ahead Lanczos is recommended [10, 13, 21]. When $\Phi_j = \mathcal{O}(1)$, we expect our approximations to be as accurate as the data warrants.

3. The viability of the two-sided Gram–Schmidt process. This section studies the central problem of how to maintain adequate duality between the two sequences of Lanczos vectors $\{p_1^*, \ldots, p_j^*\}$ and $\{q_1, \ldots, q_j\}$ at a reasonable expense. It will help to recall the corresponding technical problem in the symmetric case when $p_i = q_i$, $i = 1, \ldots, j$. See [23, 20, 28, 29]. Suppose that the three-term recurrence, in finite precision arithmetic, returns a unit vector q'_{j+1} that is not orthogonal to the previous q_i; i.e., $Q_j^* q'_{j+1}$ is not negligible. The Gram–Schmidt process replaces q'_{j+1} by a normalized version of $(I_j - Q_j Q_j^*)q'_{j+1}$. This procedure is appropriate if $Q_j^* Q_j = I_j$, but can actually make things worse if Q_j's columns are not orthogonal. The interesting question here is how much $\|I_j - Q_j^* Q_j\|_2$ can be permitted to grow and yet guarantee that $(I_j - Q_j Q_j^*)q'_{j+1}$ is orthogonal to range(Q_j) to within working accuracy.

Our problem is similar but more complicated. The formal two-sided Gram–Schmidt operator is $I_j - Q_j\Omega_j^{-1}P_j^*$, where $\Omega_j = \mathrm{diag}(P_j^* Q_j)$. How large can we permit $\|I_j - P_j^* Q_j \Omega_j^{-1}\|_2$ to grow and yet get what we want by applying $I_j - Q_j\Omega_j^{-1}P_j^*$ to $(p'_{j+1})^*$ and q'_{j+1}?

The answer in the symmetric case is that semiorthogonality defined in equation (3.1) suffices: if

$$\|Q_i^* q_{i+1}\|_1 \leq \sqrt{\epsilon} \tag{3.1}$$

holds for $i = 1, \ldots, j-1$ and if $\|Q_j^* q'_{j+1}\|_1 \leq \sqrt{\epsilon}$ then we may take $q_{j+1} = q'_{j+1}$.

We shall give a similar condition in the two-sided case–semiduality suffices. However, the definition of semiduality is not as simple as in the symmetric case, and we postpone it until more notation has been developed. An added benefit of maintaining semiduality is that the computed tridiagonal–diagonal pair of Lanczos matrices (T_j, Ω_j) is equivalent, to within round-off error, to the true "projection" of B, namely, $(P_j^* B Q_j, P_j^* Q_j)$. More precisely, we will show that the no-breakdown condition

$$\min_{i \leq j} |\omega_i| \geq (n + 10j)\epsilon \tag{3.2}$$

and the semiduality condition

$$\max_{i \leq j-1} (\|(p'_{i+1})^* Q_i |\Omega_i|^{-1/2}\|_\infty, \||\Omega_i|^{-1/2} P_i^* q'_{i+1}\|_1) \leq \sqrt{\epsilon} \tag{3.3}$$

suffice to ensure the preservation of (T_j, Ω_j) described in the previous sentence (see Theorem 5.2). It is necessary to strengthen condition (3.3) somewhat to guarantee the viability of two-sided Gram–Schmidt (GS) when (3.2) is nearly an equality. Note that in the symmetric case $\Omega_j = I_j$ and (3.3) reduces to (3.1) as claimed.

The superscript $'$ in p'_{i+1} and q'_{i+1} indicates that these are the candidate Lanczos vectors computed by the three-term recurrence, but not necessarily the actual $i+1$th Lanczos vectors. The vectors p'_{i+1} and q'_{i+1} have been normalized.

3.1. Analysis of GS. We are going to derive a sequence of matrices $\{M_j\}_{j>0}$ whose norm is the "right" factor by which duality is enhanced in GS. Recall from section 2.1 that at the end of step j the Lanczos algorithm has computed dual matrices of Lanczos vectors P_j^* and Q_j, candidate Lanczos vectors $(p'_{j+1})^*$ and q'_{j+1}, and ω_{j+1} denotes the computed value of the inner product $(p'_{j+1})^* q'_{j+1}$. We assume $\omega_{j+1} \neq 0$. Due to the loss of duality, $P_j^* Q_j \neq \Omega_j \equiv \mathrm{diag} P_j^* Q_j$ and off-diagonal entries of $P_j^* Q_j$ could be as large as 1 if the three-term recurrence is not modified.

Suppose that $\|P_j^* q'_{j+1}\|_2$ and $\|(p'_{j+1})^* Q_j\|_2$ are too big (criterion to be discussed later). GS yields new candidates $(\breve{p}_{j+1})^*$ and $\breve{q}_{j+1}$ satisfying

$$(\breve{p}_{j+1})^* = (p'_{j+1})^* (I_j - Q_j \Omega_j^{-1} P_j^*)$$

and

$$\breve{q}_{j+1} = (I_j - Q_j \Omega_j^{-1} P_j^*) q'_{j+1}.$$

Now we examine the new duality situation:

$$\begin{aligned}(\breve{p}_{j+1})^* Q_j &= (p'_{j+1})^* (I_j - Q_j \Omega_j^{-1} P_j^*) Q_j \\ &= (p'_{j+1})^* Q_j (I_j - \Omega_j^{-1} P_j^* Q_j)\end{aligned}$$

and

$$\begin{aligned}P_j^* \breve{q}_{j+1} &= P_j^* (I_j - Q_j \Omega_j^{-1} P_j^*) q'_{j+1} \\ &= (I_j - P_j^* Q_j \Omega_j^{-1}) P_j^* q'_{j+1}.\end{aligned}$$

The factor Ω_j^{-1} in the middle is alarming because we expect some ω_i to become quite small and we fear that off-diagonal entries may rise too close to 1. This feature is absent in the symmetric case. However, the situation is better than it appears.

We can obtain a more balanced expression for the duality of the new vectors $(\breve{p}_{j+1})^*$ and $\breve{q}_{j+1}$ by writing

$$\Omega_j = |\Omega_j|^{1/2}\text{sign}(\Omega_j)|\Omega_j|^{1/2}.$$

Then modified expressions for the duality, namely,

$$\begin{aligned}\breve{p}_{j+1}^* Q_j |\Omega_j|^{-1/2} &= p_{j+1}^{'*} Q_j (I_j - \Omega_j^{-1} P_j^* Q_j)|\Omega_j|^{-1/2} \\ &= p_{j+1}^{'*} Q_j |\Omega_j|^{-1/2}\text{sign}(\Omega_j^*)(\text{sign}(\Omega_j) - |\Omega_j|^{-1/2} P_j^* Q_j |\Omega_j|^{-1/2}) \end{aligned} \tag{1}$$

and

$$\begin{aligned}&|\Omega_j|^{-1/2} P_j^* \breve{q}_{j+1} \\ &= (\text{sign}(\Omega_j) - |\Omega_j|^{-1/2} P_j^* Q_j |\Omega_j|^{-1/2})\text{sign}(\Omega_j^*)|\Omega_j|^{-1/2} P_j^* q'_{j+1},\end{aligned} \tag{3.4}$$

show that the balanced "reducing factor" after applying GS is $\|M_j\|$, where

$$M_j = \text{sign}(\Omega_j) - |\Omega_j|^{-1/2} P_j^* Q_j |\Omega_j|^{-1/2}. \tag{3.5}$$

We get no benefit from the cost of GS unless $\|M_j\|$ is much less than 1.

We choose to measure duality using

$$\||\Omega_j|^{-1/2} P_j^* q_{j+1}\|_1 \quad \text{and} \quad \|p_{j+1}^* Q_j |\Omega_j|^{-1/2}\|_\infty$$

and define the effectiveness of GS using the balanced *connection* matrix M_j. Note that $\|x\|_1 = \|x^*\|_\infty$.

To illustrate the advantage of a balanced connection matrix, we applied LanFRB to the matrix B that arises from the finite difference discretization (five-point stencil) of the partial differential operator

$$L[u](\mathbf{x}) = -\Delta u + 50\nabla\cdot(u\mathbf{x}) - 125u$$

on a regular 31×31 grid over the unit square with zero boundary values [35]. Though the eigenvalue problem for B is ill posed (because the coefficients 50 and 125 are enormous compared to the grid size), this example is relevant because breakdown occurs at step 56, and at the previous step

$$|\omega_{55}| \approx 1e-13 \quad \text{and} \quad \|\Omega_{55}^{-1} T_{55}\|_1 \approx 11\|B\|_1.$$

Figures 1 and 2 display the absolute values of the entries of the unbalanced connection matrix $I_{55} - \Omega_{55}^{-1} P_{55}^* Q_{55}$ and the balanced connection matrix M_{55} on a semilog scale. Though $\|I_{55} - \Omega_{55}^{-1} P_{55}^* Q_{55}\|_1 \approx 1e-4$, the norm of the balanced operator is much less, $\|M_{55}\|_1 \approx 2e-11$.

Recall that $(\breve{p}_{j+1})^*$ and $\breve{q}_{j+1}$ are obtained from $(p'_{j+1})^*$ and q'_{j+1} by GS. If

$$\|(p'_{j+1})^* Q_j |\Omega_j|^{-1/2}\|_\infty \le \frac{\epsilon}{\|M_j\|_\infty}, \tag{3.6}$$

then, by (3.4), one trivially has

$$\|(\breve{p}_{j+1})^* Q_j |\Omega_j|^{-1/2}\|_\infty \le \epsilon.$$

Similarly, if

$$\||\Omega_j|^{-1/2} P_j^* q'_{j+1}\|_1 \le \frac{\epsilon}{\|M_j\|_1}, \tag{3.7}$$

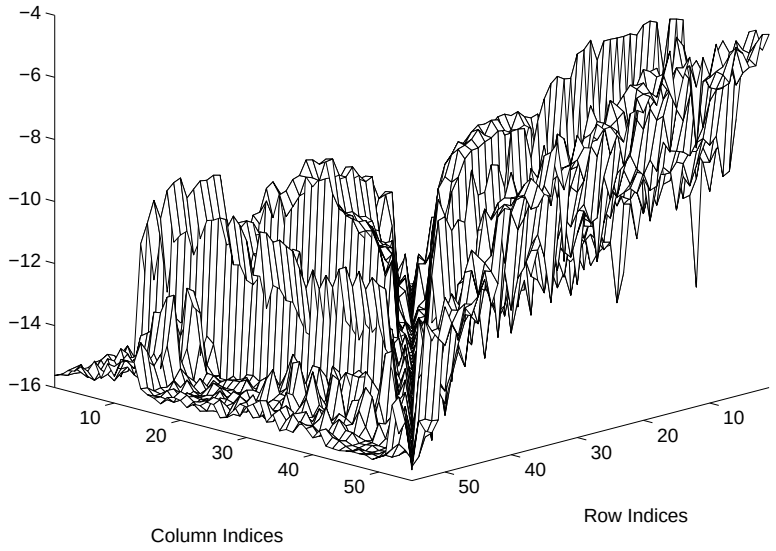


FIG. 1. $I_{55} - \Omega_{55}^{-1} P_{55}^* Q_{55}$ *(unbalanced op.);* $\|I_{55} - \Omega_{55}^{-1} P_{55}^* Q_{55}\|_1 \approx 1\mathrm{e}-4$.

then, by (3.4),

$$\||\Omega_j|^{-1/2} P_j^* \breve{q}_{j+1}\|_1 \le \epsilon.$$

The sequence

$$(\max(\|(p'_{j+1})^* Q_j |\Omega_j|^{-1/2}\|_\infty, \||\Omega_j|^{-1/2} P_j^* q'_{j+1}\|_1))_{j\ge 1} \tag{3.8}$$

tends to increase with j gradually until the last term is too big. At that step a correction step is made (that is, $q'_{j+1} \to \breve{q}_{j+1}$ and $p'_{j+1} \to \breve{p}_{j+1}$). This change reduces the latest term in (3.8) to ϵ. Hence semilog graphs of (3.8) look sawtoothed.

We use this perspective on GS to find the "right" definition of semiduality. For overall efficiency we want to minimize the total number of corrections and particularly avoid unnecessary corrections near the end of a Lanczos run. So we seek the weakest conditions that give adequate levels of duality. To this end we explicitly ensure that

$$\max(\|(p'_{j+1})^* Q_j |\Omega_j|^{-1/2}\|_\infty \|M_j\|_\infty, \|M_j\|_1 \||\Omega_j|^{-1/2} P_j^* q'_{j+1}\|_1) \le \epsilon. \tag{3.9}$$

Condition (3.9) takes no account of ω_{j+1}, and if $|\omega_{j+1}|$ is too small it is essential not to accept $(p'_{j+1})^*$ and q'_{j+1}. So, in addition to (3.9) we must take account of possible growth in $\|M_{j+1}\|_1$. To analyze this, note that the nonzero part of the rightmost column of M_{j+1} is

$$\Omega_j^{-1/2} P_j^* q'_{j+1} |\omega_{j+1}|^{-1/2}.$$

Since

$$\|\Omega_j^{-1/2} P_j^* q'_{j+1} |\omega_{j+1}|^{-1/2}\|_1 \le \|M_{j+1}\|_1,$$

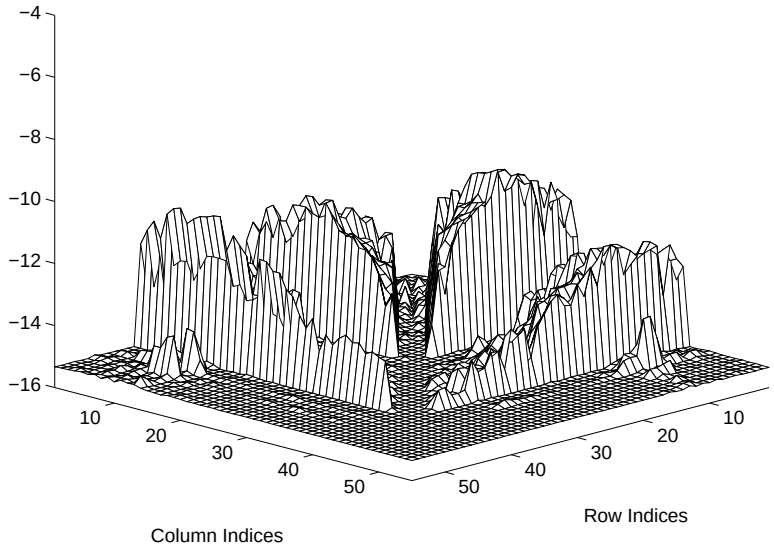


FIG. 2. M_{55} *(balanced op.);* $\|M_{55}\|_1 \approx 2\mathrm{e}-11$.

a necessary condition for (3.9) to hold at step $j+2$ without correction is that

$$\frac{\||\Omega_j|^{-1/2}P_j^* q'_{j+1}\|_1}{|\omega_{j+1}|^{1/2}}\||\Omega_{j+1}|^{-1/2}P_{j+1}^* q'_{j+2}\|_1 \leq \|M_{j+1}\|_1\||\Omega_{j+1}|^{-1/2}P_{j+1}^* q'_{j+2}\|_1 \leq \epsilon. \tag{3.10}$$

The square root of (3.10) yields

$$(\||\Omega_j|^{-1/2}P_j^* q'_{j+1}\|_1\||\Omega_{j+1}|^{-1/2}P_{j+1}^* q'_{j+2}\|_1)^{1/2} \leq \epsilon^{1/2}|\omega_{j+1}|^{1/4},$$

and this is guaranteed by

$$\max(\||\Omega_j|^{-1/2}P_j^* q'_{j+1}\|_1, \||\Omega_{j+1}|^{-1/2}P_{j+1}^* q'_{j+2}\|_1) \leq \epsilon^{1/2}|\omega_{j+1}|^{1/4}. \tag{3.11}$$

A similar argument applied to $(p_{j+1}^*)'$ yields our definition of semiduality.

DEFINITION 3.1. *Semiduality holds at step* $j+1$ *if for* $i \leq j$,

$$\max(\|p_{i+1}^{'*}Q_i|\Omega_i|^{-1/2}\|_\infty, \||\Omega_i|^{-1/2}P_i^* q'_{i+1}\|_1) \leq \epsilon^{1/2}\ |\omega_{i+1}|^{1/4}.$$

4. The Lanczos algorithm with correction. In this section, we present a practical and efficient implementation of the Lanczos algorithm with correction (LanCor hereafter). Several implementation details are addressed and relevant properties of the computed quantities are established. Extensive work from [7] on how to implement the Lanczos recurrences is summarized in section 4.1. In section 4.2 we discuss how to correct the duality loss and what effect this has on the computed quantities. An efficient implementation of correction steps called retroactive correction is discussed and justified in section 4.3. In section 4.4 we will show how to compute $(p'_{j+1})^* Q_j$ and $P_j^* q'_{j+1}$ without accessing P_j^* and Q_j and using only $\mathcal{O}(j)$ floating point operations and storage per step. LanCor is given in section 4.5.

4.1. Implementing the three-term recurrences. A prerequisite to the analyses of later sections is an understanding of how nearly dual consecutive pairs of left and right Lanczos vectors can be. We say that *local duality* holds at step j if

$$\max_{1<i\le j}(|p_i^* q_{i-1}|, |p_{i-1}^* q_i|) \le 4\epsilon. \tag{4.1}$$

In [7] it was proved that local duality is maintained to within a (theoretically necessary but generally unrealistic) factor of n by the implementation of the three-term recurrences below called LanLD. LanLD stands for the Lanczos algorithm maintaining local duality.

ALGORITHM (LANLD).

Start: $p_1 = p/\|p\|_2, q_1 = q/\|q\|_2, \omega_0 = 1, \beta_1 = \gamma_1 = 0, \omega_1 = p_1^* q_1$.

Iterate: For i=1,MaxStep

1. $r_i^* = p_i^* B - \frac{\gamma_i \omega_i}{\omega_{i-1}} p_{i-1}^*, \quad s_i = Bq_i - q_{i-1}\frac{\beta_i \omega_i}{\omega_{i-1}}$
2. $\alpha_i \in \{p_i^* s_i, r_i^* q_i\}$
3. $r_i^* := r_i^* - \frac{\alpha_i}{\omega_i} p_i^*, \quad s_i := s_i - q_i \frac{\alpha_i}{\omega_i}$
4. $\alpha_i^l = r_i^* q_i, \qquad \alpha_i^r = p_i^* s_i$ /* these are small corrections to α_i */
5. $r_i^* := r_i^* - \frac{\alpha_i^l}{\omega_i} p_i^*, \quad s_i := s_i - q_i \frac{\alpha_i^r}{\omega_i}$
6. $\beta_{i+1} = \|r_i^*\|_2, \quad \gamma_{i+1} = \|s_i\|_2$
7. Check for invariant subspace. See equations (4.3) and (4.4).
8. $p_{i+1}^* = r_i^*/\beta_{i+1}, \quad q_{i+1} = s_i/\gamma_{i+1}$
9. $\omega_{i+1} = p_{i+1}^* q_{i+1}$
10. Check for breakdown: $|\omega_{i+1}| < (n + 10(i+1))\epsilon$
11. Check for convergence periodically (see section 2.2)

Remark 1. The meaning of step 2 is that α_i can be assigned either value $p_i^* s_i$ or $r_i^* q_i$; it does not matter which.

Remark 2. Local duality is maintained by steps 4 and 5.

The properties of the quantities computed by LanLD are summarized as follows. We assume that the no-breakdown condition holds,

$$\min_{1\le i\le j} |\omega_i| \ge (n + 10j)\epsilon, \tag{4.2}$$

and that the no-invariant subspace conditions hold: for $1 \le i \le j$,

$$\beta_{i+1} \ge \max(\sqrt{\epsilon}(\Phi_i + 1 + \psi)\|B\|_2, |\alpha_i^l/\omega_i|) \tag{4.3}$$

and

$$\gamma_{i+1} \ge \max(\sqrt{\epsilon}(\Phi_i + 1 + \psi)\|B\|_2, |\alpha_i^r/\omega_i|). \tag{4.4}$$

Here Φ_i is the growth factor defined by (2.7), and the constant ψ accounts for the rounding error introduced when the operator B is applied.

In this section, we add two more error bounds to our model of the computed quantities which are proved to be realistic in [7]. First, the Lanczos vectors satisfy the perturbed three-term recurrences

$$\beta_{j+1} p_{j+1}^* = p_j^* B - \frac{\alpha_j + \alpha_j^l}{\omega_j} p_j^* - \frac{\gamma_j \omega_j}{\omega_{j-1}} p_{j-1}^* - f_j^* \tag{4.5}$$

and

$$q_{j+1}\gamma_{j+1} = Bq_j - q_j \frac{\alpha_j + \alpha_j^r}{\omega_j} - q_{j-1}\frac{\beta_j \omega_j}{\omega_{j-1}} - g_j, \tag{4.6}$$

where the matrices $F_j = [f_1, \ldots, f_j]$ and $G_j = [g_1, \ldots, g_j]$ are such that

$$\max(\|F_j\|_2, \|G_j\|_2) \leq \epsilon(\Phi_j + 1)\|B\|_2. \tag{4.7}$$

Second, the tiny refinements to the trailing bits of each α_i to maintain local duality,

$$D_j^l = \mathrm{diag}(\alpha_i^l)_{i=1}^j \quad \text{and} \quad D_j^r = \mathrm{diag}(\alpha_i^r)_{i=1}^j, \tag{4.8}$$

satisfy

$$\max(\|D_j^l\|_2, \|D_j^r\|_2) \leq \epsilon(\Phi_j + 1)\|B\|_2. \tag{4.9}$$

4.2. Properties of the computed quantities. Recall from section 3 that the loss of duality of the candidate Lanczos vectors to the previous Lanczos vectors is corrected using a version of the GS process. Our model of the properties of the quantities computed by LanCor is obtained by amending the model for LanLD to account for correction steps.

To correct the loss of duality of the $(i+1)$st Lanczos vectors to the previous Lanczos vectors at the end of a Lanczos step we first compute

$$x_i^* = (p_{i+1}')^* Q_{i-1}, \quad y_i = P_{i-1}^* q_{i+1}^{'}. \tag{4.10}$$

Next we "re-biorthogonalize" or correct the candidate Lanczos vectors:

$$(\breve{p}_{j+1})^* = (p_{j+1}')^* - x_i^* \Omega_{i-1}^{-1} P_{i-1}^* \tag{4.11}$$

and

$$\breve{q}_{j+1} = q_{j+1}' - Q_{i-1}\Omega_{i-1}^{-1} y_i. \tag{4.12}$$

Let $\mathcal{I}_j$ denote the set of all indices i up to and including j at which correction steps are taken, let e_i denote the ith column of the $j \times j$ identity matrix, and let

$$\Lambda_j = D_j^l + \sum_{i \in \mathcal{I}_j} \beta_{i+1} e_i (x_i^*, 0), \quad \Upsilon_j = D_j^r + \sum_{i \in \mathcal{I}_j} \begin{bmatrix} y_i \\ 0 \end{bmatrix} e_i^* \gamma_{i+1}. \tag{4.13}$$

Recall that D_j^l and D_j^r are defined in equation (4.8).

For the purpose of illustration, $T_{40} + \Upsilon_{40}$ corresponding to a model problem discussed in [25] is displayed on a semilog scale in Figure 3.

The governing equations for LanCor are

$$P_j^* B = (T_j + \Lambda_j)\Omega_j^{-1} P_j^* + e_j \beta_{j+1} p_{j+1}^* + F_j^* \tag{4.14}$$

and

$$BQ_j = Q_j \Omega_j^{-1}(T_j + \Upsilon_j) + q_{j+1}\gamma_{j+1} e_j^* + G_j. \tag{4.15}$$

The matrices Υ_j (for upper) and Λ_j (for lower) are upper and lower triangular matrices of spikes, one spike for each correction step. Local duality, (4.1), (4.7), and

$$\max(\|\Lambda_j \Omega_j^{-1/2}\|_2, \|\Omega_j^{-1/2}\Upsilon_j\|_2) \leq \sqrt{\epsilon}(\Phi_j + 1)\|B\|_2 \tag{4.16}$$

are also realistic for LanCor [7].

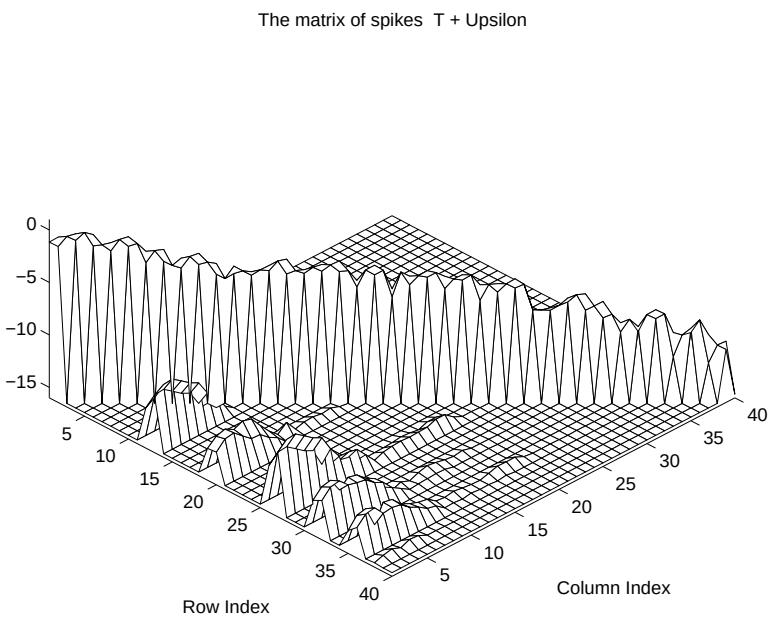


FIG. 3. $T_{40} + \Upsilon_{40}$.

4.3. Retroactive correction. In this section we show how to implement a correction step,

$$p'_{j+1} \to \breve{p}_{j+1}, \quad q'_{j+1} \to \breve{q}_{j+1}.$$

As in the symmetric case, correction steps are taken in pairs,

$$p'_{j+1} \to \breve{p}_{j+1}, \quad q'_{j+1} \to \breve{q}_{j+1},$$

$$p'_{j} \to \breve{p}_{j}, \quad q'_{j} \to \breve{q}_{j}.$$

The reason for correcting the jth Lanczos vectors with the $j+1$th is the same as in the symmetric case, and for the convenience of the reader we revisit the explanation. In practice the loss of duality is gradual,

$$\|P^*_{j-1} q_j\| \approx \|P^*_{j-1} q_{j+1}\|,$$

and thus $\|P^*_{j-1} q_j\|$ is less than but approximately equal to the semiduality threshold. Consider the $j+2$th Lanczos vector

$$q_{j+2}\gamma_{j+2} = Bq_{j+1} - q_{j+1}\frac{\alpha_{j+1}}{\omega_{j+1}} - q_j \frac{\beta_{j+1}\omega_{j+1}}{\omega_j}.$$

Multiply by P^*_{j-1} to obtain

$$P^*_{j-1} q_{j+2}\gamma_{j+2} = P^*_{j-1} Bq_{j+1} - P^*_{j-1} q_{j+1}\frac{\alpha_{j+1}}{\omega_{j+1}} - P^*_{j-1} q_j \frac{\beta_{j+1}\omega_{j+1}}{\omega_j}.$$

If the jth Lanczos vectors are not corrected along with the $j+1$th Lanczos vectors, then

$$\|P_{j-1}^* q_{j+2}\|\gamma_{j+2} \approx \|P_{j-1}^* q_j\| \frac{\beta_{j+1}|\omega_{j+1}|}{|\omega_j|},$$

which is often just below the semiduality threshold. Correcting the jth Lanczos vectors with the $j+1$th substantially reduces $\|P_{j-1}^* q_{j+2}\|$, and this postpones the next correction step.

Correction steps are implemented so that each Lanczos vector is transferred from slow storage to fast storage and back again only once. Following [24], we call this *retroactive* correction. That is, correcting the jth Lanczos vectors with the $j+1$th doubles the number of floating point operations per correction step, but the amount of data transfer is the same. Retroactive correction by the two-sided modified GS algorithm is implemented as follows.

ALGORITHM (RETROACTIVE CORRECTION).

Iterate: For $i = 1 \ldots j-1$,

1. $p_{j+1} := p_{j+1} - p_i(\omega_i^{-*}(q_i^* p_{j+1}))$
2. $p_j := p_j - p_i(\omega_i^{-*}(q_i^* p_j))$
3. $q_{j+1} := q_{j+1} - q_i(\omega_i^{-1}(p_i^* q_{j+1}))$
4. $q_j := q_j - q_i(\omega_i^{-1}(p_i^* q_j))$

Recover local duality

1. $p_{j+1} := p_{j+1} - p_j(\omega_j^{-*}(q_j^* p_{j+1}))$
2. $q_{j+1} := q_{j+1} - q_j(\omega_j^{-1}(p_j^* q_{j+1}))$

Retroactive correction changes the relations among the computed quantities. Nonetheless, a careful analysis shows that as long as semiduality is maintained properties (4.1), (4.7), and (4.16) are also realistic for LanCor with retroactive correction [7]. It is important *not* to normalize p_j and q_j after a retroactive correction step. The reason is that normalizing p_j and q_j in this case nonnegligibly alters β_j and γ_j. This needlessly complicates the approximate three-term recurrences among the computed quantities.

4.4. Monitoring the loss of duality. We must correct for the loss of duality when either $p_{j+1}^* Q_j |\Omega_j^{-1/2}|$ or $|\Omega_j^{-1/2}| P_j^* q_{j+1}$ increases to $\sqrt{\epsilon}|\omega_{j+1}^{1/4}|$. To compute these vectors at each step is about as costly as correcting the loss of duality at each step. The same problem arises in the symmetric case. Compromise symmetric Lanczos algorithms avoid this costly step by updating a recurrence estimating the loss of orthogonality at each step [11, 19, 20, 24, 28, 29]. In this section we extend the partial reorthogonalization (PRO) algorithm from the symmetric Lanczos algorithm [28, 29]. Recurrence relations for $p_{j+1}^* Q_j$ and $P_j^* q_{j+1}$ for the unnormalized two-sided Lanczos algorithm based on a model of the properties of the quantities computed in finite precision arithmetic appear in [36].

The sequence of vectors $(p_{i+1}^* Q_i)$ and $(P_i^* q_{i+1})$ satisfy a three-term recurrence which we now derive. Let $\omega_{i,j} = p_i^* q_j$. In this notation, $\omega_i = \omega_{i,i}$.

Suppose that a correction step is not taken at step j (i.e., in computing the $(j+1)$st Lanczos vectors). Multiplying equation (4.5) by Q_j and substituting BQ_j according to (4.15), we have

$$\beta_{j+1} p_{j+1}^* Q_j = p_j^* Q_j \left[\Omega_j^{-1}(T_j + \Upsilon_j) - \frac{\alpha_j + \alpha_j^l}{\omega_j} I \right]$$

$$-\frac{\gamma_j\omega_j}{\omega_{j-1}}p_{j-1}^*Q_j+\omega_{j,j+1}\gamma_{j+1}e_j^*+p_j^*G_j-f_j^*Q_j. \tag{4.17}$$

Similarly multiplying (4.6) by P_j^* on the left and substituting in P_j^*B according to (4.14), we have

$$P_j^*q_{j+1}\gamma_{j+1}=\left[(T_j+\Lambda_j)\Omega_j^{-1}-\frac{\alpha_j+\alpha_j^r}{\omega_j}I\right]P_j^*q_j$$

$$-\frac{\beta_j\omega_j}{\omega_{j-1}}P_j^*q_{j-1}+e_j\beta_{j+1}\omega_{j+1,j}+F_j^*q_j-P_j^*g_j. \tag{4.18}$$

To further reduce these equations, we first need to discuss some additional relations among the computed quantities. First we show that the correction terms

$$p_j^*Q_j\Omega_j^{-1}\Upsilon_j \quad\text{and}\quad \Lambda_j\Omega_j^{-1}P_j^*q_j$$

negligibly affect the loss of duality among the computed Lanczos vectors. For this reason, these matrices are not used to estimate the loss of duality and are not stored. Since j is not a correction step, equation (4.13) implies that

$$\Lambda_j\Omega_j^{-1}p_j^*q_je_j=D_j^le_j=e_j\alpha_j^l.$$

Substitute the definition of semiduality, (3.3), and (4.16) to find

$$\left\|\Lambda_j\Omega_j^{-1}\begin{bmatrix}P_{j-1}^*q_j\\0\end{bmatrix}\right\|_2\le\|\Lambda_j\Omega_j^{-1/2}\|_2\|\Omega_{j-1}^{-1/2}P_{j-1}^*q_j\|_2\le(\Phi_j+1)\epsilon\|B\|_2.$$

The analysis of $p_j^*Q_j\Omega_j^{-1}\Upsilon_j$ is similar.

Next we expand terms such as $P_j^*q_j$:

$$p_j^*Q_j=(p_j^*Q_{j-1},0)+\omega_je_j^* \quad\text{and}\quad P_j^*q_j=\begin{bmatrix}P_{j-1}^*q_j\\0\end{bmatrix}+\omega_je_j. \tag{4.19}$$

By equation (4.19) and the definition of T_j, we have

$$\begin{aligned}&p_j^*Q_j\left(\Omega_j^{-1}T_j-\frac{\alpha_j+\alpha_j^l}{\omega_j}I\right)\\&\quad=(p_j^*Q_{j-1},0)\left[\Omega_j^{-1}T_j-\frac{\alpha_j+\alpha_j^l}{\omega_j}I\right]+\gamma_j\omega_je_{j-1}^t-\alpha_j^le_j^t\end{aligned} \tag{4.20}$$

and

$$\begin{aligned}&\left(T_j\Omega_j^{-1}-\frac{\alpha_j+\alpha_j^r}{\omega_j}I\right)P_j^*q_j\\&\quad=\left[T_j\Omega_j^{-1}-\frac{\alpha_j+\alpha_j^r}{\omega_j}I\right]\begin{bmatrix}P_{j-1}^*q_j\\0\end{bmatrix}+e_{j-1}\beta_j\omega_j-e_j\alpha_j^r.\end{aligned} \tag{4.21}$$

Since step j is not a correction step, the last row of Υ_j and the last column of Λ_j are zero. That is why they do not appear.

We also need the identities

$$p_{j-1}^* Q_j = (p_{j-1}^* Q_{j-2}, 0, 0) + \omega_{j-1} e_{j-1}^* + \omega_{j-1,j} e_j^* \tag{4.22}$$

and

$$P_j^* q_{j-1} = \begin{bmatrix} P_{j-2}^* q_{j-1} \\ 0 \\ 0 \end{bmatrix} + \omega_{j-1} e_{j-1} + \omega_{j,j-1} e_j. \tag{4.23}$$

Finally, substitute equations (4.22) and (4.20) into equation (4.17) and equations (4.23) and (4.21) into equation (4.18) and the desired recurrences appear:

$$\begin{aligned} \beta_{j+1} p_{j+1}^* Q_j = {} & (p_j^* Q_{j-1}, 0) \left(\Omega_j^{-1} T_j - \frac{\alpha_j}{\omega_j} I \right) \\ & - \frac{\gamma_j \omega_j}{\omega_{j-1}} ((p_{j-1}^* Q_{j-2}, 0, 0) + \omega_{j-1,j} e_j^t) \\ & + (\omega_{j,j+1} \gamma_{j+1} + \alpha_j^r - \alpha_j^l) e_j^t + \mathcal{O}(\epsilon(\Phi_j + 1)\|B\|_2) \end{aligned} \tag{4.24}$$

and

$$\begin{aligned} P_j^* q_{j+1} \gamma_{j+1} = {} & \left(T_j \Omega_j^{-1} - \frac{\alpha_j}{\omega_j} I \right) \begin{bmatrix} P_{j-1}^* q_j \\ 0 \end{bmatrix} \\ & - \frac{\beta_j \omega_j}{\omega_{j-1}} \left[\begin{bmatrix} P_{j-2}^* q_{j-1} \\ 0 \\ 0 \end{bmatrix} + \omega_{j,j-1} e_j \right] \\ & + e_j (\beta_{j+1} \omega_{j+1,j} + \alpha_j^l - \alpha_j^r) + \mathcal{O}(\epsilon(\Phi_j + 1)\|B\|_2). \end{aligned} \tag{4.25}$$

Note that certain computable terms such as $\omega_{j,j-1}$ and α_j^l which are $\mathcal{O}(\epsilon(\Phi_j + 1)\|B\|_2)$ are not included in $\mathcal{O}(\epsilon(\Phi_j+1)\|B\|_2)$. This is done because these computable quantities are used to estimate $P_j^* q_{j+1}$ and $p_{j+1}^* Q_j$ in the next section.

4.4.1. An implementation of the monitoring algorithm. LanCor is similar to LanLD, but with additional work done (if necessary) between LanLD iterations to maintain semiduality. The perturbed recurrences (4.24) and (4.25) are invoked to compute $\beta_{j+1} p_{j+1}^* Q_j$ and $P_j^* q_{j+1} \gamma_{j+1}$ after β_{j+1} and γ_{j+1} are computed as in step 6 of LanLD. The decision whether or not to correct duality loss is then made as determined in section 3.1. In this section we show how to implement the recurrences to obtain accurate estimates of the duality loss.

In LanCor at each Lanczos step j the candidate $j+1$th Lanczos vectors are explicitly "dualized" against the $j-1$th Lanczos vectors (extended local duality) and then the jth Lanczos vectors (local duality).

$P_j^* q_{j+1}$ is estimated by h_{j+1}. Initially h_2 and h_3 are exact, and for $j > 2$,

$$h_{j+1} \gamma_{j+1} = \left(T_j \Omega_j^{-1} - \frac{\alpha_j}{\omega_j} I \right) \begin{bmatrix} h_j \\ 0 \end{bmatrix} - \begin{bmatrix} h_{j-1} \\ 0 \\ 0 \end{bmatrix} \frac{\beta_j \omega_j}{\omega_{j-1}} - e_{j-2} \alpha_j^r. \tag{4.26}$$

To account for the perturbation of the three-term recurrences by correction steps, $\epsilon \mathrm{diag}|\Omega_j^{-1} T_j|$ is added to the right-hand side above if the loss of the duality of q_{j-1} and q_j was corrected. The $j-1$ and j entries of the estimate h_{j+1} are assigned the exact values $p_{j-1}^* q_{j+1}$ and $p_j^* q_{j+1}$. Maintaining extended local duality sweeps the α_j^r term from the jth entry to the $j-2$th entry, hence the $e_{j-2} \alpha_j^r$ above. The estimate of $p_{j+1}^* Q_j$ is computed by the similar recurrence.

4.5. The implementation of LanCor. In this section we summarize the implementation of LanCor, the Lanczos algorithm maintaining semiduality.

ALGORITHM (LanCor).

Start: $p_1 = p/\|p\|_2, q_1 = q/\|q\|_2, \omega_0 = 1, \beta_1 = \gamma_1 = 0, \omega_1 = p_1^* q_1$.

Iterate: For i=1,MaxStep

1. $r_i^* = p_i^* B - \frac{\gamma_i \omega_i}{\omega_{i-1}} p_{i-1}^*, \quad s_i = Bq_i - q_{i-1}\frac{\beta_i \omega_i}{\omega_{i-1}}$
2. $\alpha_i = r_i^* q_i$
3. $r_i^* = r_i^* - \frac{\alpha_i}{\omega_i} p_i^*, \quad s_i = s_i - q_i \frac{\alpha_i}{\omega_i}$
4. Maintain extended local duality (see section 4.4.1)
5. $\alpha_i^l = r_i^* q_i, \qquad \alpha_i^r = p_i^* s_i$
6. $r_i^* := r_i^* - \frac{\alpha_i^l}{\omega_i} p_i^*, \quad s_i := s_i - q_i \frac{\alpha_i^r}{\omega_i}$
7. $\beta_{i+1} = \|r_i^*\|_2, \quad \gamma_{i+1} = \|s_i\|_2$
8. Check for invariant subspace. See equations (4.3) and (4.4).
9. $p_{i+1}^* = r_i^*/\beta_{i+1}, \quad q_{i+1} = s_i/\gamma_{i+1}$
10. $\omega_{i+1} = p_{i+1}^* q_{i+1}$
11. Check for breakdown: $|\omega_{i+1}| < (n + 10(i+1))\epsilon$
12. Monitor duality loss (see section 4.4.1)
13. Correct duality loss only if necessary (see section 4.3)
14. Check for convergence after a correction step only (see section 2.2)

Remark 3. The loss of duality among the computed Lanczos vectors corresponds to either a near breakdown of the algorithm or the convergence of a Ritz value to an eigenvalue of B [1]. For this reason it is more efficient to check for convergence only after correction steps.

5. Preserved quantities. LanCor applied to an operator B after j successful steps yields matrices P_j^* and Q_j of Lanczos vectors and the reduced tridiagonal–diagonal pencil (T_j, Ω_j). In this section we compare the computed quantities to the corresponding exact quantities determined by B, the row span of P_j^*, and the column span of Q_j. We say that a computed quantity is preserved when it is as close to the corresponding exact quantity as the data warrants.

Our main result is that semiduality suffices to preserve (T_j, Ω_j). See section 5.3.

The analysis is more complicated than in the symmetric case. We must avoid perturbations that are proportional to $\|\Omega_j^{-1}\|_2 = 1/\min_{1\le i\le j} |\omega_i|$.

5.1. Exact projections. The operator

$$\Pi_j = Q_j (P_j^* Q_j)^{-1} P_j^* \tag{5.1}$$

is the oblique projection corresponding to the computed Lanczos vectors. The projection of B onto the spaces spanned by the Lanczos vectors is the matrix $\Pi_j B \Pi_j$. See Definition 2.1. The Lanczos vectors are dual if and only if $P_j^* Q_j$ is diagonal and nonsingular.

We recover exactly dual bases corresponding to the computed Lanczos vectors by use of the LDU factorization of $P_j^* Q_j$:

$$P_j^* Q_j = L_j \hat{\Omega}_j U_j. \tag{5.2}$$

Recall that Ω_j =diag$(P_j^* Q_j)$. If $\hat{\Omega}_j$ is nonsingular, then substitute (5.2) into (5.1) to obtain

$$\Pi_j = Q_j U_j^{-1}\ \hat{\Omega}_j^{-1}\ L_j^{-1} P_j^*. \tag{5.3}$$

The rows of

$$\hat{P}_j^* = L_j^{-1} P_j^* \tag{5.4}$$

and the columns of

$$\hat{Q}_j = Q_j U_j^{-1} \tag{5.5}$$

are dual since

$$\hat{P}_j^* \hat{Q}_j = L_j^{-1} P_j^* Q_j U_j^{-1} = \hat{\Omega}_j.$$

Two-sided GS applied to P_j^* and Q_j yields $\hat{P}_j^*$ and $\hat{Q}_j$. Next define $\hat{T}_j$ by

$$\hat{T}_j = \hat{P}_j^* B \hat{Q}_j. \tag{5.6}$$

Note that $\hat{T}_j$ is *not* tridiagonal. The representation of $\Pi_j B \Pi_j$ with respect to the bases $\hat{\Omega}_j^{-1}\hat{P}_j^*$ and $\hat{Q}_j$ is $\hat{\Omega}_j^{-1}\hat{T}_j$. This matrix is equivalent to the pencil $(\hat{T}_j, \hat{\Omega}_j)$. This pencil is analogous to the orthogonal projection of the operator onto the span of the computed Lanczos vectors in the symmetric case.

5.2. Conditions for preservation. Each computed Lanczos vector is the sum of the vector which exactly satisfies a three-term recurrence and another vector whose norm is proportional to the machine epsilon ϵ. See equations (4.5) and (4.6). This result is typical of Krylov subspace methods [20]. This perturbation of the Lanczos vectors causes perturbations of the diagonal elements of Ω by approximately ϵ and perturbations of the tridiagonal elements of T by approximately $\epsilon\|B\|$. We show that exactly correcting the loss of duality among the computed Lanczos vectors does not change the pencil (T, Ω) by significantly more than these amounts. We call this property the preservation of T and Ω. The elements of T and Ω are not in general determined to working (or full relative) precision. This implies that $T\Omega^{-1}$ and $\Omega^{-1}T$ are not determined to full absolute precision.

This section addresses the problem of determining necessary and sufficient conditions for three properties of the computed quantities to hold. The three properties are (1) that $W = P^*Q$ admits an LDU factorization, (2) that the diagonal matrix D is approximately Ω, and (3) that L and U are well conditioned. To be precise, we determine realistic sufficient conditions for any complex $n \times n$ matrix W with nonzero diagonal elements to admit an LDU factorization

$$W = LDU \tag{5.7}$$

such that

$$\|\mathrm{diag}(W) - D\|_2 \leq 2\epsilon \tag{5.8}$$

and

$$\max(\|L^{-1}\|_2, \|U^{-1}\|_2) < 2. \tag{5.9}$$

In our case $W = P^*Q$. By equation (5.2) $D = \hat{\Omega}$ holds and (5.8) immediately implies the preservation of Ω. The preservation of T is discussed in section 5.3.

Our results are given in the two theorems below. Theorem 5.1 gives necessary and sufficient conditions for (5.7) and (5.8) to hold. Theorem 5.2 gives sufficient

conditions for all three properties to hold which are only slightly stronger than the hypotheses of Theorem 5.1 (i.e., the lower bound on $|\omega_j|$ increases from $2j\epsilon$ to $10j\epsilon$). We ultimately increase the latter lower bound by $n\epsilon$ to $(n+10j)\epsilon$ to account for the discrepancy between the computed and exact values of $P_j^* Q_j$.

For any matrix C let $triu'(C)$ denote its strictly upper triangular part.

THEOREM 5.1. *Let W be a $j \times j$ complex matrix and let $\Omega = \mathrm{diag}(W) = \mathrm{diag}(\omega_i)$. Suppose that $\epsilon > 0$, $j > 2$, $(j-2)\epsilon < 1$, and W satisfies the following hypotheses:*

$$\min_{1 \le i \le j} |\omega_i| \ge 2(j-2)\epsilon, \tag{5.10}$$

$$\max(\|\Omega^{-1/2} triu'(W)\|_1, \|\Omega^{-1/2} triu'(W^*)\|_1) \le \sqrt{\epsilon}. \tag{5.11}$$

Then equations (5.7) *and* (5.8) *hold.*

Proof. See [7].

Remark 4. The second hypothesis (5.11) is equivalent to the semiduality condition (3.3). Note that the factor of $|\omega_{i+1}|^{1/4}$ that appears in Definition 3.1 to maintain the viability of the two-sided GS process is not necessary to ensure the preservation of (T, Ω) in Theorems 5.1 and 5.2.

One approach to proving Theorem 5.1 is to apply the perturbation theory for Gaussian elimination. Many papers have recently appeared on this subject [2, 33, 34]. To guarantee condition (5.8), all of these general perturbation bounds require significantly stronger hypotheses than Theorem 5.1.

Theorem 5.1 gives necessary and sufficient conditions for the preservation of Ω (i.e., condition (5.8)). By increasing the lower bound on $|\omega_j|$ by a factor of 5, we will show that the computed quantities are preserved. Due to the difficulty of this task, we must be satisfied with unachievable but realistic bounds.

THEOREM 5.2. *Let W be a $j \times j$ complex matrix and let $\Omega = \mathrm{diag}(W) = \mathrm{diag}(\omega_i)$. Suppose that $\epsilon > 0$, $j > 2$, $(j-2)\epsilon < 1$, and Ω is nonsingular. Suppose in addition that W satisfies the hypothesis* (5.11)*, and that*

$$\min |\omega_i| \ge 10j\epsilon. \tag{5.12}$$

Then equations 5.7 *to* 5.9 *hold.*

Proof. See [7].

The idea of the proofs is to decompose W into the sequence of extensions

$$W_{i+1} = \begin{pmatrix} W_i & y_i \\ x_i^t & \omega_{i+1} \end{pmatrix}. \tag{5.13}$$

We define the sequence $\{\kappa_i\}_{i=1}^j$ corresponding to a norm $\|.\|$ by

$$\max(\|x_i\|, \|y_i\|) = \kappa_i. \tag{5.14}$$

As in [17], we then extract the worst-case information corresponding to $\{\kappa_i\}_{i=1}^j$.

5.3. Preservation of T. The pencil computed by the three-term recurrences can eventually become of larger order than the original matrix and clearly differs from the one that would be produced in exact arithmetic, but this can never happen if semiduality holds. In this section, we show that if semiduality is maintained then the computed pencil (T_j, Ω_j) agrees with the exact oblique projection of the spans of the computed left and right Lanczos vectors to full absolute precision.

Correcting the loss of duality of the $(j+1)$st Lanczos vectors to the previous Lanczos vectors replaces the vectors computed by the three-term recursion with (approximations of) $\hat{p}_{j+1}$ and $\hat{q}_{j+1}$, where $\hat{p}_{j+1}$ denotes column $j+1$ of $\hat{P}_k$ and $\hat{q}_{j+1}$ denotes column $j+1$ of $\hat{Q}_k$. The corresponding elements of T and Ω change to (approximations of) the corresponding elements of the dense matrix $\hat{T}$ and the diagonal matrix $\hat{\Omega}$. We want to know how large this perturbation is and in particular when it is negligible.

The following theorem established the preservation of T for LanCor. Recall that Theorem 5.1 establishes the preservation of Ω. The proof uses the properties of the computed quantities established in section 4 and this section. The hypotheses of Theorem 5.2 are the no-breakdown and no-invariance conditions from section 4.1 and the semiduality condition; for $1 \le i \le j$,

$$\max(\||\Omega_i|^{-1/2} P_i^* q_{i+1}\|_1, \||\Omega_i|^{-1/2} Q_i^* p_{i+1}\|_1) \le \sqrt{\epsilon}. \tag{5.15}$$

THEOREM 5.3. *Let B be a complex $n \times n$ matrix and let P_j^* and Q_j be the matrices of Lanczos vectors computed by LanCor. Let T_j denote the computed tridiagonal and let $\hat{T}_j$ be defined as in equation* (5.6)*. If $\Omega_j = \mathrm{diag}(P_j^* Q_j)$ satisfies the no-breakdown condition* (4.2) *the no-invariant subspace conditions* (4.3) *and* (4.4) *and semiduality holds* (*see* (5.15))*, then for Φ_j defined in equation* (2.7)

$$\|\hat{T}_j - T_j\|_2 = \mathcal{O}(j(\Phi_j + 1)\epsilon \|B\|_2)$$

holds.

Proof. See [7].

6. Numerical experiments. The Lanczos algorithm maintaining local duality only (LanLD), semiduality (LanCor), and full duality (LanFRB) have been applied to many tasks. We present the results for one representative example here. All computations were done in MATLAB on an IBM Power Workstation with machine precision $\epsilon \approx 2 \;\; 10^{-16} = 2\mathrm{e}-16$.

6.1. The Tolosa matrix. We illustrate the properties of LanCor using the Tolosa matrix A of order $n = 2000$ from the Harwell Boeing sparse matrix collection. The computational task is to compute the largest eigenvalues of A to half precision. We choose to compute the 50 largest eigenvalues because this emphasizes the difference between LanLD and LanCor. A has 5184 nonzero entries and $\|A\|_1 \approx 1\mathrm{e}+6.8$. Since A averages less than three nonzeros per row, the inner products in the three-term recurrences cost nearly as much as the matrix–vector multiplications in terms of floating point operations.

The eigenvalue problem for A is known to be ill conditioned and A is known to possess multiple eigenvalues [26]. For theoretical purposes, we computed the eigenvalues of A by the QR algorithm and observed that the spectral radius of A is approximately $1\mathrm{e}+3.4$. For this reason, the MATLAB function `balance`() was applied to A to obtain a balanced matrix B diagonally similar to A. For this B $\|B\|_1 \approx \|B\|_\infty \approx 1\mathrm{e}+4.0$ holds. Since $\|B\|_2^2 \le \|B\|_1 \|B\|_\infty$, balancing yields a matrix whose Euclidean norm is within a factor of 4 of its spectral radius. QR applied to B computes three real eigenvalues— -12.098, -24.196, and -36.294—of multiplicity 382; the remaining eigenvalues are distinct and well separated. Though the eigenvalues of A and B are the same (barring underflow), the computed eigenvalues of A and B by QR (without

TABLE 1

Correction step	238	264	283	295	310	323	333	347	363
No. eigenvalues	14	22	26	28	34	38	42	48	56

balancing) agree to from full to half relative precision. We will compare the eigenvalues of B computed by the QR algorithm to the three implementations of the Lanczos algorithm (LanFRB, LanCor, and LanLD) and Arnoldi's method.

We did not compare the Lanczos algorithm to the implicitly restarted Arnoldi iteration [30, 15]. Arnoldi's method does establish a lower bound for the number of steps required by implicitly restarted Arnoldi iteration. Implicit restarts can be incorporated into the Lanczos algorithm as well as Arnoldi's method [12].

6.2. Results. All three implementations of the Lanczos algorithm computed the requested 50 eigenvalues to the same high accuracy. For Arnoldi's method, LanFRB, and LanLD the Ritz values are checked every 50 iterations. In each case $p_1 = q_1$ is the same random vector (normal distribution). LanFRB and LanCor have identical convergence properties; this is a consequence of the preservation of the pencil (T, Ω) (see section 5). The reward for maintaining semiduality is that fewer Lanczos steps are required to complete the given task. In this case, LanFRB and LanCor required 400 and 363 Lanczos steps, respectively, while LanLD and Arnoldi's method required 450 and 400, respectively. Because convergence is checked after correction steps instead of periodically in LanCOR, fewer Lanczos steps are required than for LanFRB in this experiment.

No copies of converged eigenvalues appear among the Ritz values when semi- or full duality is maintained, but copies do appear among the Ritz values computed by LanLD.

LanCor takes 16 correction steps to compute the requested eigenvalues, 1/25th as many as LanFRB. Table 1 gives the last 9 steps at which the duality loss is corrected in LanCor and the number of converged Ritz values at that step.

For comparison we applied Arnoldi's method with modified GS orthogonalization to this task and computed eigenvalues of B to the same accuracy [27]. The number of converged Ritz values near the end of the run are displayed in Table 2.

TABLE 2

Arnoldi step	50	100	150	200	250	300	350	400
No. eigenvalues	0	0	0	6	16	30	48	70

In this example Arnoldi's method, LanFRB, and LanCor yield approximate eigenvalues of similar accuracy for a fixed number of steps; the differences in the number of steps is due to the convergence criteria.

LanFRB, LanCor, and LanLD were each stable in the sense that

$$\|\Omega^{-1}T\|_1 \approx 35\|B\|_2,$$

even though $\min|\omega_i| \approx 2\mathrm{e}-5$ (see section 2.3).

The $\log_{10}$ of the number of floating point operations (flops) in these Lanczos and Arnoldi runs are tabulated in Table 3. The flop count for applying the operator (a sparse matrix–vector multiplication in this case) is given in column OP; column EIG displays the flop count for solving the reduced eigenvalue problems by

TABLE 3
Results for balanced Tolosa.

FLOPS ($\log_{10}$)	OP	EIG	(BI-) ORTH	ALGO	TOTAL
Arnoldi (400)	6.6	9.5	8.8	6.9	9.6
LanFRB (400)	7.0	7.7	**9.1**	7.7	9.1
LanCor (363)	7.0	7.8	**8.1**	7.8	8.4
LanLD (450)	7.0	7.8	None	8.0	8.2

the QR algorithm for Arnoldi's method and by the differential QD algorithm, an algorithm that exploits the tridiagonal structure for the Lanczos-based procedures [8]. (BI-)ORTH gives the flop count for maintaining the duality or orthogonality of the basis vectors, and column ALGO contains the remaining flop count. The number of steps required by each method is given in parenthesis below the algorithm name.

We were surprised at the large number of flops required by the QR algorithm in Arnoldi's method in this example. For comparison note that computing the eigenvalues of B by the QR algorithm requires $1\mathrm{e}+11$ flops.

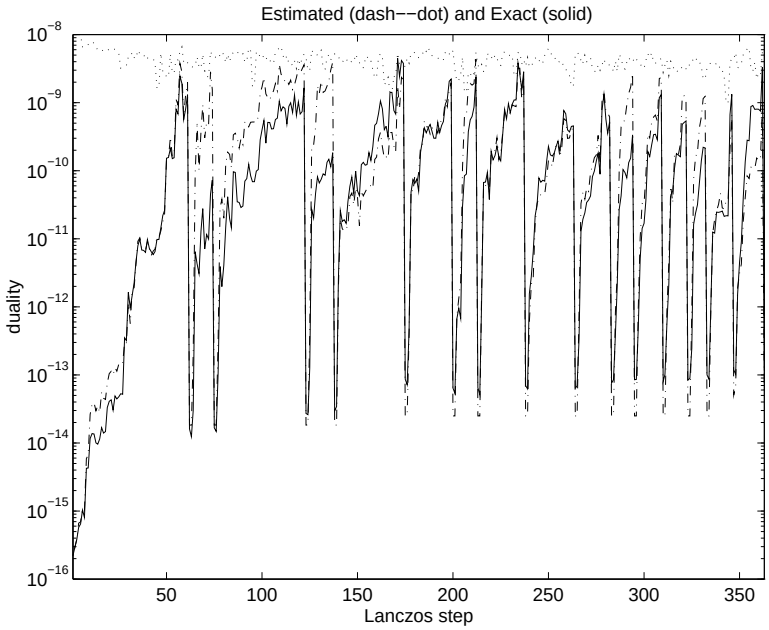


FIG. 4. *Numerical duality for LanCor applied to Tolosa matrix.*

Maintaining semiduality requires an order of magnitude fewer flops than full duality. Also, LanCor has an order of magnitude fewer flops than Arnoldi's method.

LanLD requires the fewest flops and takes the most steps. The low flop count is due to the less rigorous stopping criteria. The error introduced by accepting a Ritz value as an eigenvalue is estimated by the minimum of three quantities: the

distance from the Ritz value to the nearest remaining Ritz value and the left and the right unnormalized residuals. Recall from section 2.2 that if v is an eigenvector of $\Omega^{-1}T$, then Qv is used to approximate an eigenvector of B, and reliable accuracy estimates must factor in the shrinkage $\|Qv\|_2/\|v\|_2$. We observed shrinkage, i.e., $\|Qv\|_2/\|v\|_2 \approx .01$, for all the Ritz vectors of interest in this example. For this reason the error estimates based on unnormalized Ritz vectors are 100 times too small. In LanLD the Lanczos vectors are not stored and so the shrinkage of the Ritz vectors is not available. Even if the Lanczos vectors are stored, forming the eigenvectors requires 1e+8.6 real floating point operations (see section 2.2). That is, if we demand reliability from LanLD similar to that of LanCor, the LanLD flop count will increase above the LanCor flop count.

We conclude by illustrating the effectiveness of the duality-monitoring algorithm of section 4.4.1. Figure 4 compares the $\log_{10}$ of our estimate of

$$\max(\||\Omega_i|^{-1/2}P_i^* q_{i+1}\|_1, \||\Omega_i|^{-1/2}Q_i^* p_{i+1}\|_1) \tag{2}$$

at each step (dash–dot line) i to the exact value (solid line). Each spike indicates a correction step. The dotted line across the top of the figure is the target threshold $\epsilon^{1/2}|\omega_i|^{1/4}$.

REFERENCES

[1] Z. Bai (1992), *Error analysis of the Lanczos algorithm for the nonsymmetric eigenvalue problem*, Math. Comp., 62, pp. 209–226.

[2] A. Barrlund (1991), *Perturbation bounds for the LDL^H and LU decompositions*, BIT, 31, pp. 358–363.

[3] Z. Bai, D. Day, and Q. Ye (1995), *ABLE: An Adaptive Block Lanczos Method of the Eigenvalue Problem*, Technical report 95-07, Mathematics Department, University of Kentucky, Lexington, KY.

[4] D. Boley and G. Golub (1990), *The nonsymmetric Lanczos algorithm and controllability*, Systems Control Lett., 16, pp. 97–105.

[5] J. Cullum and R. Willoughby (1985), *Lanczos Algorithms for Large Symmetric Eigenvalue Computations*, Birkhäuser Boston, Cambridge, MA.

[6] J. Cullum, W. Kerner, and R. Willoughby (1989), *A generalized nonsymmetric Lanczos procedure*, Comp. Phys. Comm., 53, pp. 19–48.

[7] D. Day (1993), *Semi-Duality in the Two-Sided Lanczos Algorithm*, Ph.D. thesis, University of California, Berkeley, CA.

[8] D. Day (1995), *The differential QD algorithm for the tridiagonal eigenvalue problem*, manuscript.

[9] T. Ericsson and A. Ruhe (1980), *The spectral transformation Lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems*, Math. Comp., 35, pp. 1251–1268.

[10] R. Freund, M. Gutknecht, and N. Nachtigal (1993), *An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices*, SIAM J. Sci. Stat. Comput., 14, pp. 137–158.

[11] R. Grimes, J. Lewis, and H. Simon (1994), *A shifted block Lanczos algorithm for solving generalized eigenproblems*, SIAM J. Matrix Anal. Appl., 15, pp. 228–272.

[12] E. Grimme, D. Sorensen, and P. Van Dooren (1994), *Model Reduction of State Space Systems via an Implicitly Restarted Lanczos Method*, TR94-21, Rice University, Houston, TX.

[13] M. Gutknecht (1992), *A completed theory of the unsymmetric Lanczos process and related algorithms*, SIAM J. Matrix Anal. Appl., Part I, 13, pp. 594–639, Part II, 15, pp. 15–58.

[14] W. Kahan, B. Parlett, and E. Jiang (1982), *Residual bounds on approximate eigensystems of nonnormal matrices*, SIAM J. Numer. Anal., 19, pp. 470–484.

[15] R. Lehoucq (1995), *Analysis and Implementation of an Implicitly Restarted Arnoldi Iteration*, TR95-13, Rice University, Houston, TX.

[16] C. Lanczos (1950), *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Nat. Bur. Stand., 45, pp. 255–282.

[17] A. Ostrowski (1937), *Über die determinanten mit über wiegender hauptdiagonale*, Commet. Math. Helv., 10, pp. 69–96 (1937) or *Alexander Ostrowski Collected Mathematical Papers* 1, Birkhaüser-Verlag, pp. 31–59 (1983).
[18] B. Parlett (1974), *The Rayleigh quotient algorithm iteration and some generalizations for nonnormal matrices*, Math. Comp., 28, pp. 679–693.
[19] B. Parlett and D. Scott (1979), *The Lanczos algorithm with selective orthogonalization*, Math. Comp., 33, pp. 217–238.
[20] B. Parlett (1980), *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ.
[21] B. Parlett, D. Taylor, and Z.-S. Liu (1985), *A look-ahead Lanczos algorithm for unsymmetric matrices*, Math. Comp., 44, pp. 105–124.
[22] B. Parlett (1992), *Reduction to tridiagonal form and minimal realizations*, SIAM J. Matrix Anal. Appl., 13, pp. 567–593.
[23] B. Parlett (1992), *The rewards for maintaining semi-orthogonality among Lanczos vectors*, J. Numer. Linear Algebra Appl., 1, pp. 243–267.
[24] B. Parlett, B. Nour Omid, and Z.-S. Liu (1985), *How to Maintain Semi-Orthogonality Among Lanczos Vectors*, CPAM-420, Center for Pure and Applied Mathematics, University of California, Berkeley, CA.
[25] A. Ruhe (1993), *The two-sided Arnoldi algorithm for nonsymmetric eigenvalue problems*, in Matrix Pencils, LNM 973, B. Kågström and A. Ruhe, eds., Springer-Verlag, Berlin, Heidelberg, New York, pp. 104–120.
[26] A. Ruhe (1995), *Rational Krylov, A Practical Algorithm for Large Sparse Nonsymmetric Matrix Pencils*, Report UCB/CSD-95-871, Computer Science Division (EECS), University of California, Berkeley, CA.
[27] Y. Saad (1980), *Variations of Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebra Appl., 34, pp. 269–295.
[28] H. Simon (1984), *Analysis of the symmetric Lanczos algorithm with reorthogonalization*, Linear Algebra Appl., 61, pp. 101–131.
[29] H. Simon (1984), *The Lanczos algorithm with partial reorthogonalization*, Math. Comp., 42, pp. 115–142.
[30] D. Sorensen (1992), *Implicit application of polynomial filters in a k-step Arnoldi process*, SIAM J. Matrix Anal. Appl., 13, pp. 357–385.
[31] G. Stewart and J. Sun (1990), *Matrix Perturbation Theory*, Academic Press, New York.
[32] G. Stewart (1993), *On the perturbation of LU, Cholesky, and QR factorizations*, SIAM J. Matrix Anal. Appl., 14, pp. 1141–1145.
[33] J. G. Sun (1991), *Perturbation bounds for the Cholesky and QR factors*, BIT, 31, pp. 341–352.
[34] J. G. Sun (1992), *Component-wise perturbation bounds for some matrix decompositions*, BIT, 32, pp. 702–714.
[35] C. Tong and Q. Ye (1995), *Analysis of the Finite Precision Bi-Conjugate Gradient Algorithm for Nonsymmetric Linear Systems*, preprint.
[36] H. I. van der Veen and K. Vuik (1995), *Bi-Lanczos with partial orthogonalization*, Computers and Structures, 56, pp. 605–613.
[37] J. Wilkinson (1965), *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, UK.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 3, pp. 590–614, July 1997

005

ON COMPUTING STABLE LAGRANGIAN SUBSPACES OF HAMILTONIAN MATRICES AND SYMPLECTIC PENCILS*

WEN-WEI LIN† AND CHERN-SHUH WANG‡

Abstract. This paper presents algorithms for computing stable Lagrangian invariant subspaces of a Hamiltonian matrix and a symplectic pencil, respectively, having purely imaginary and unimodular eigenvalues. The problems often arise in solving continuous- or discrete-time H^∞-optimal control, linear-quadratic control and filtering theory, etc. The main approach of our algorithms is to determine an isotropic Jordan subbasis corresponding to purely imaginary (unimodular) eigenvalues by using the associated Jordan basis of the square of the Hamiltonian matrix (the $S + S^{-1}$-transformation of the symplectic pencil). The algorithms preserve structures and are numerically efficient and reliable in that they employ only orthogonal transformations in the continuous case.

Key words. stable Lagrangian subspace, purely imaginary eigenvalue, Hamiltonian matrix, unimodular eigenvalue, symplectic pencil

AMS subject classifications. 47A15, 15A18, 15A21,, 15A22, 15A24

PII. S0895479894272712

1. Introduction.

A matrix $M \in \mathbf{R}^{2n\times 2n}$ is said to be Hamiltonian if $JM = (JM)^T$, where $J \equiv J_n = \left[\begin{smallmatrix} O_n & I_n \\ -I_n & O_n \end{smallmatrix}\right]$. Here I_n is the $n \times n$ identity matrix and O_n is the $n \times n$ zero matrix. A matrix $S \in \mathbf{R}^{2n\times 2n}$ is symplectic if $S^T JS = J$. A linear pencil $N - \lambda L$ with $N, L \in \mathbf{R}^{2n\times 2n}$ is said to be symplectic if $NJN^T = LJL^T$. If we partition a Hamiltonian matrix M and a symplectic pencil $N - \lambda L$ comfortably with J, respectively, then we have

$$M = \begin{bmatrix} A & G \\ H & -A^T \end{bmatrix}, \quad G = G^T, \quad H = H^T, \tag{1.1}$$

and

$$N = \begin{bmatrix} A & O \\ -H & I \end{bmatrix}, \quad L = \begin{bmatrix} I & G \\ O & A^T \end{bmatrix}, \quad G = G^T, \quad H = H^T. \tag{1.2}$$

Our interest in the Hamiltonian matrix M in (1.1) and the symplectic pencil $N - \lambda L$ in (1.2), respectively, stems from the fact that if

$$\begin{bmatrix} A & G \\ H & -A^T \end{bmatrix} \begin{bmatrix} \Omega_1 \\ \Omega_2 \end{bmatrix} = \begin{bmatrix} \Omega_1 \\ \Omega_2 \end{bmatrix} W, \quad \Omega_1,\ \Omega_2,\ W \in \mathbf{R}^{n\times n}, \tag{1.3}$$

then $X = -\Omega_2\Omega_1^{-1}$ (if Ω_1^{-1} exists) solves the continuous-time algebraic Riccati equation (CARE)

$$-XGX + XA + A^T X + H = 0, \tag{1.4}$$

* Received by the editors August 10, 1994; accepted for publication (in revised form) by P. Van Dooren July 12, 1996.
http://www.siam.org/journals/simax/18-3/27271.html
† Institute of Applied Mathematics, Tsing Hua University, Hsinchu, Taiwan (wwlin@am.nthu.edu.tw).
‡ Department of Applied Mathematics, Chiao Tung University, Hsinchu, Taiwan (d788101@am.nthu.edu.tw).

and if

$$\begin{bmatrix} A & O \\ -H & I \end{bmatrix} \begin{bmatrix} \Omega_1 \\ \Omega_2 \end{bmatrix} = \begin{bmatrix} I & G \\ O & A^T \end{bmatrix} \begin{bmatrix} \Omega_1 \\ \Omega_2 \end{bmatrix} W, \tag{1.5}$$

then $X = -\Omega_2\Omega_1^{-1}$ (if Ω_1^{-1} exists) solves the discrete-time algebraic Riccati equation (DARE)

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

In fact, the Hamiltonian matrix and the symplectic pencil are often derived from continuous- and discrete-time optimal control problems, respectively, e.g., [5, 6, 8, 10, 11, 13, 14]. To obtain an optimizer, especially a stabilizing optimizer, of optimal control problems, one must compute a particular invariant subspace satisfying (1.3) or (1.5). This particular invariant subspace is usually referred to as a stable Lagrangian subspace.

DEFINITION 1.1. *A subspace $\mathcal{S} \subset \mathbf{R}^{2n}$ is isotropic if*

$$x^T J y = 0 \quad \text{for all } x, y \in \mathcal{S}.$$

DEFINITION 1.2. *A subspace $\mathcal{Y} \subset \mathbf{R}^{2n}$ is called an M-stable isotropic subspace if $\mathcal{Y}$ satisfies that* (i) $M\mathcal{Y} \subset \mathcal{Y}$, (ii) $\mathcal{Y}$ *is isotropic, and* (iii) $\mathrm{Re}(\lambda(M|_{\mathcal{Y}})) \le 0$. *Here $\lambda(M|_{\mathcal{Y}})$ denotes an eigenvalue of M restricted in $\mathcal{Y}$.*

DEFINITION 1.3. *A subspace $\mathcal{W} \subset \mathbf{R}^{2n}$ is called an (N, L)-stable isotropic subspace if* (i) $\mathcal{W}$ *is invariant under (N, L)* [25]; *i.e., there is a subspace $\mathcal{V}$ such that $N\mathcal{W}$, $L\mathcal{W} \subset \mathcal{V}$*; (ii) $\mathcal{W}$ *is isotropic*; *and* (iii) $|\lambda((N, L)|_{\mathcal{W}})| \le 1$.

DEFINITION 1.4. *If $\mathcal{Y}_{\mathcal{L}} \subset \mathbf{R}^{2n}$ is an M-stable isotropic subspace with $\dim(\mathcal{Y}_{\mathcal{L}}) = n$, then $\mathcal{Y}_{\mathcal{L}}$ is called an M-stable Lagrangian subspace.*

DEFINITION 1.5. *If $\mathcal{W}_{\mathcal{L}} \subset \mathbf{R}^{2n}$ is an (N, L)-stable isotropic subspace with $\dim(\mathcal{W}_{\mathcal{L}}) = n$, then $\mathcal{W}_{\mathcal{L}}$ is called an (N, L)-stable Lagrangian subspace.*

For the continuous-time case, it is known that an M-stable Lagrangian subspace is closely related to an internally stabilizing controller of an H^∞-control system [5, 8]. In linear-quadratic control problems in which (A, G) is stabilizable with G positive semidefinite, we can obtain the unique "weak" stabilizing symmetric solution of CARE (1.4), and therefore an optimal controller by computing the unique M-stable Lagrangian subspace [14, 28]. In addition, several applications in Wiener filtering theory [26] and network synthesis [1] also need to compute an M-stable Lagrangian subspace. This is the reason why we are interested in computing an M-stable Lagrangian subspace. Unfortunately, an M-stable Lagrangian subspace does not always exist, while some nonzero purely imaginary eigenvalues of M have odd partial multiplicities. A counterexample can be found in [21].

To guarantee the existence of an M-stable Lagrangian subspace, M must satisfy the following assumption.

(A1) *The partial multiplicities of all purely imaginary eigenvalues are all even.*

If we require that

(R1) *the M-stable Lagrangian subspace $\mathcal{Y}_{\mathcal{L}}$ have the lowest Jordan degree (that is, there is no other M-stable Lagrangian subspace having total Jordan degree smaller than that of $\mathcal{Y}_{\mathcal{L}}$), then the desired M-stable Lagrangian subspace $\mathcal{Y}_{\mathcal{L}}$ is unique determined.*

We will discuss the details of this result in the next section.

The first purpose of this paper is to propose an efficient, reliable, and structure-preserving algorithm for computing the M-stable Lagrangian subspace satisfying (R1) under the assumption (A1). For Hamiltonian matrices with purely imaginary eigenvalues, Clements and Glover [5] proposed an eigenvector deflation technique that guarantees that the eigenvalues appear with the correct pairing. This is certainly an advantage over the general QR or QZ method [12, 15, 24], but this method still ignores the structure in part during the process. In another recent paper, Ammar and Mehrmann [2] proposed an elegant method, only using symplectic orthogonal transformations to compute the M-stable Lagrangian subspace. Combining the method with at least one step of defect correction is highly advisable. But, there are still numerical difficulties in convergence of deflation steps if purely imaginary eigenvalues occur [20, section 18, p. 143].

To avoid the numerical difficulties mentioned above, we shall develop a stable and structure-preserving algorithm as a preprocessing step to deflate all purely imaginary eigenvalues and to get a reduced Hamiltonian matrix having no purely imaginary eigenvalues. Then the rest of the M-stable Lagrangian subspace corresponding to stable eigenvalues with negative real parts can be computed by some reliable algorithms, such as in [2, 23, 29]. In our algorithm, we first compute the skew-Hamiltonian Schur decomposition of M^2 by using the numerically stable square reduced algorithm of Van Loan [27]. Then, we apply the algorithm proposed in [3] or [17] to the skew-Hamiltonian Schur matrix to determine the Jordan subbasis corresponding to the nonpositive eigenvalues of M^2. These algorithms are numerically reliable and need only $O(n^2)$ flops if the number of nonpositive eigenvalues of M^2 is of order $O(1)$. Based on elementary linear algebra theory, we can determine an associated Jordan subbasis Y corresponding to purely imaginary eigenvalues of M by using the Jordan subbasis corresponding to nonpositive eigenvalues of M^2. Under the assumption (A1) that each purely imaginary eigenvalue has even partial multiplicities, by applying an isotropicity requirement, we can separate an isotropic Jordan subbasis Υ corresponding to each first half of Jordan blocks of purely imaginary eigenvalues from Y. Indeed, the subspace span$\{\Upsilon\}$ lies on the M-stable Lagrangian subspace. Consequently, we deflate the isotropic subbasis Υ from M by using symplectic orthogonal transformations and get a reduced Hamiltonian matrix having no purely imaginary eigenvalues.

For the discrete-time case, an (N, L)-stable Lagrangian subspace also play an important role for H^∞-optimal or linear-quadratic control problems. In linear-quadratic control problems in which (A, G) is stabilizable with G positive semidefinite, the unique "weak" stabilizing symmetric solution of DARE (1.6) can be obtained by computing the (N, L)-stable Lagrangian subspace [13]. For the H^∞-control problem a detailed treatment of the suboptimal controller versus the H^∞-optimal control is not available. The suboptimal case is treated in detail in [10, 11]. Although a factorization theory similar to [5] has not been developed for the discrete-time case, we still consider computing the (N, L)-stable Lagrangian subspace of $N - \lambda L$ from a theoretical point of view. To ensure the existence and uniqueness of the desired (N, L)-stable Lagrangian subspace with lowest Jordan degree, a related assumption and requirement as in the continuous-time case are listed as follows.

(A2) *The partial multiplicities of all unimodular eigenvalues of $N - \lambda L$ are even.*

(R2) *The (N, L)-stable Lagrangian subspace $\mathcal{W}_\mathcal{L}$ has the lowest Jordan degree. (That is, there is no other (N, L)-stable Lagrangian subspace having total Jordan degree smaller than that of $\mathcal{W}_\mathcal{L}$.)*

As in a continuous-time case, we can also develop a reliable and structure-preserving algorithm as a preprocessing step to deflate all unimodular eigenvalues and get a reduced symplectic pencil having no unimodular eigenvalues. Then the rest of the (N, L)-stable Lagrangian subspace can be computed by algorithms of [19] or [29]. In our algorithm we consider the $S+S^{-1}$-transformation of the symplectic pencil $N-\lambda L$ [18], i.e.,

$$\Gamma - \lambda\Delta \equiv \left[\left(NJL^T + LJN^T\right) - \lambda LJL^T\right] J^T, \tag{1.7}$$

and then we compute the skew-Hamiltonian Schur pencil form of $\Gamma - \lambda\Delta$ by using the numerically stable algorithm proposed in [22]. As in the continuous-time case, we first compute a Jordan subbasis corresponding to eigenvalues of $\Gamma - \lambda\Delta$ with magnitudes between -2 and 2 by algorithms of [3] or [17] and then use it to determine an isotropic Jordan subbasis corresponding to each first half of Jordan blocks of unimodular eigenvalues of $N - \lambda L$. Further, we deflate this subbasis of $N - \lambda L$ by symplectic transformations and get a reduced symplectic pencil having no unimodular eigenvalues.

For convenience, we list some notation which are adopted in this paper.

Z_p denotes an orthonormal matrix which forms an orthonormal subbasis of M^2 corresponding to the zero eigenvalue with the Jordan degree of p; i.e., for any nonzero vector $v \in \text{span}\{Z_p\}$,

$$(M^2)^p v = 0 \ \text{ and } \ (M^2)^{p-1} \neq 0.$$

$\widetilde{Z}_p$ denotes the matrix $[Z_1, \ldots, Z_p]$.

Y_p denotes an orthonormal matrix which forms an orthonormal subbasis of M corresponding to the zero eigenvalue with the Jordan degree of p; i.e., for any nonzero vector $v \in \text{span}\{Y_p\}$,

$$M^p v = 0 \ \text{ and } \ M^{p-1} \neq 0.$$

$\widetilde{Y}_p$ denotes the matrix $[Y_1, \ldots, Y_p]$.

Υ_s denotes an orthonormal matrix which forms an orthonormal subbasis of the maximal M-stable isotropic subspace corresponding to each first half of Jordan blocks of zero eigenvalue.

$J^{(\ell)}(\lambda)$ denotes an $\ell \times \ell$ elementary Jordan matrix corresponding to λ; i.e.,

$$J^{(\ell)}(\lambda) = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & & 1 \\ & & & \lambda \end{bmatrix}_{\ell\times\ell}.$$

$\Lambda^{(\ell)}(0)$ denotes an $\ell \times \ell$ matrix with

$$\Lambda^{(\ell)}(0) = \left[\begin{array}{c|c} O_{\ell-1} & \\ \hline & \delta_\ell \end{array}\right], \quad \delta_\ell = 1 \ \text{ or } \ 0.$$

$e_j \equiv e_j^{(n)}$ is the jth column vector of $n \times n$ identity matrix I_n.

$\mathcal{N}(A)$ denotes the null space of matrix A.

All script (calligraphic) capital letters, e.g., $\mathcal{Y}$, $\mathcal{W}$, etc. denote vector subspaces.

This paper is organized as follows. In section 2 we summarize some preliminary results. In sections 3 and 4 we develop numerically reliable algorithms to compute the desired isotropic subspaces of a Hamiltonian matrix and a symplectic pencil, respectively, corresponding to purely imaginary and unimodular eigenvalues. In section 5, we show some numerical results to illustrate the numerical reliability of our algorithms.

2. Preliminary. In this section, we review some important properties of a real Hamiltonian matrix and a real symplectic pencil which have been developed and exploited for several years. First, we state a theorem of [16] which gives a canonical form of a Hamiltonian matrix.

THEOREM 2.1 (see [16]). *Let $M \in \mathbf{R}^{2n\times 2n}$ be a Hamiltonian matrix. Then there is a symplectic matrix $S \in \mathbf{R}^{2n\times 2n}$ such that*

$$S^{-1}MS = \left[\begin{array}{c|c} \operatorname{diag}\{J_0, 0, T_1, J_\nu^T\} & \operatorname{diag}\{\Lambda_0, E_\mu, T_2, D_\nu\} \\ \hline \operatorname{diag}\{0, E_{-\mu}, 0, -D_\nu\} & \operatorname{diag}\{-J_0^T, 0, -T_1^T, -J_\nu\} \end{array}\right], \tag{2.1}$$

where $\mu = (\mu_1, \ldots, \mu_{k_2})^T \in \mathbf{R}^{k_2}$, $T_1 \in \mathbf{R}^{k_3\times k_3}$ with $Re(\lambda(T_1)) < 0$, $\nu = (\nu_1, \ldots, \nu_{k_4})^T \in \mathbf{R}^{k_4}$, and

$$\begin{aligned} J_0 &= \operatorname{diag}\{J^{(m_1)}(0), \ldots, J^{(m_{k_1})}(0)\}, \\ \Lambda_0 &= \operatorname{diag}\{\Lambda^{(m_1)}(0), \ldots, \Lambda^{(m_{k_1})}(0)\}, \\ E_\mu &= \operatorname{diag}\{E^{(n_1)}(\mu_1), \ldots, E^{(n_{k_2})}(\mu_{k_2})\}, \\ E_{-\mu} &= \operatorname{diag}\{E^{(n_1)}(-\mu_1), \ldots, E^{(n_{k_2})}(-\mu_{k_2})\} \end{aligned}$$

with (n_j an even integer),

$$E^{(n_j)}(\mu_j) = \begin{bmatrix} 0 & & & & \mu_j \\ & & & \mu_j & 1 \\ & & \cdot & -1 & \\ & \mu_j & \cdot & & \\ \mu_j & 1 & & & 0 \end{bmatrix}_{n_j\times n_j},$$

$$E^{(n_j)}(-\mu_j) = \begin{bmatrix} 0 & & & -1 & -\mu_j \\ & & \cdot & -\mu_j & \\ & 1 & \cdot & & \\ -1 & -\mu_j & & & \\ -\mu_j & & & & 0 \end{bmatrix}_{n_j\times n_j},$$

$$\begin{aligned} J_\nu &= \operatorname{diag}\{J^{(\ell_1)}(0), \ldots, J^{(\ell_{k_4})}(0)\}, \\ D_\nu &= \operatorname{diag}\{D^{(\ell_1)}(\nu_1), \ldots, D^{(\ell_{k_4})}(\nu_{k_4})\} \end{aligned}$$

with (ℓ_j an odd integer)

$$D^{(\ell_j)}(\nu_j) = \begin{bmatrix} 0 & & & -\nu_j \\ & & \nu_j & \\ & \cdot & & \\ \cdot & & & \\ -\nu_j & & & 0 \end{bmatrix}_{\ell_j\times\ell_j},$$

and

$$n = \sum_{j=1}^{k_1} m_j + \sum_{j=1}^{k_2} n_j + k_3 + \sum_{j=1}^{k_4} \ell_j. \qquad \square$$

By Theorem 2.1, we see that the Hamiltonian matrix M contains zero eigenvalues and purely imaginary eigenvalues $\pm i\mu_j$ for $j = 1, \ldots, k_2$ and $\pm i\nu_j$ for $j = 1, \ldots, k_4$.

Under assumption (A1), the canonical form (2.1) becomes a simpler form,

$$S^{-1}MS = \left[\begin{array}{c|c} \operatorname{diag}\{J_0, 0, T_1\} & \operatorname{diag}\{\Lambda_0, E_\mu, T_2\} \\ \hline \operatorname{diag}\{0, E_{-\mu}, 0\} & \operatorname{diag}\{-J_0^T, 0, -T_1^T, \} \end{array}\right], \tag{2.2}$$

where $\mu, T_1, J_0, \Lambda_0, E_\mu, E_{-\mu}$ are given in (2.1) with $n = \sum_{j=1}^{k_1} m_j + \sum_{j=1}^{k_2} n_j + k_3$.

Partition the symplectic matrix $S = [S_1, S_2, S_3, \widehat{S}_1, \widehat{S}_2, \widehat{S}_3]$ with the block type (2.2). Furthermore, we partition

$$S_1 = [S_1^{(1)}, \ldots, S_1^{(k_1)}] \text{ and } \widehat{S}_1 = [\widehat{S}_1^{(1)}, \ldots, \widehat{S}_1^{(k_1)}]$$

comfortably with block type of J_0 and write $S_1^{(j)}$ and $\widehat{S}_1^{(j)}$ in the column vector forms

$$S_1^{(j)} = [s_1^{(1,j)}, \ldots, s_{m_j}^{(1,j)}] \text{ and } \widehat{S}_1^{(j)} = [\hat{s}_1^{(1,j)}, \ldots, \hat{s}_{m_j}^{(1,j)}]$$

for $j = 1, \ldots, k_1$. If $\delta_j = 1$ (the (m_j, m_j)th element of $\Lambda^{(m_j)}(0)$) for some $j \in \{1, \ldots, k_1\}$, then the maximal M-stable isotropic subspace with lowest Jordan degree of $\operatorname{span}\{S_1^{(j)}, \widehat{S}_1^{(j)}\}$ is

$$\mathcal{S}_1^{(j)} = \operatorname{span}\{S_1^{(j)}\}. \tag{2.3}$$

If $\delta_j = 0$ for some $j \in \{1, \ldots, k_1\}$ (here m_j must be even), then the maximal M-stable isotropic subspace with lowest Jordan degree of $\operatorname{span}\{S_1^{(j)}, \widehat{S}_1^{(j)}\}$ is

$$\mathcal{S}_1^{(j)} = \operatorname{span}\{s_1^{(1,j)}, \ldots, s_{m_j/2}^{(1,j)}, \hat{s}_{m_j/2}^{(1,j)}, \ldots, \hat{s}_{m_j}^{(1,j)}\}. \tag{2.4}$$

Partition

$$S_2 = [S_2^{(1)}, \ldots, S_2^{(k_2)}] \text{ and } \widehat{S}_2 = [\widehat{S}_2^{(1)}, \ldots, \widehat{S}_2^{(k_2)}]$$

with the block type E_μ and write $S_2^{(j)}$ and $\widehat{S}_2^{(j)}$ in the column vector forms

$$S_2^{(j)} = [s_1^{(2,j)}, \ldots, s_{n_j}^{(2,j)}] \text{ and } \widehat{S}_2^{(j)} = [\hat{s}_1^{(2,j)}, \ldots, \hat{s}_{n_j}^{(2,j)}]$$

for $j = 1, \ldots, k_2$. The maximal M-stable isotropic subspace with lowest Jordan degree of $\operatorname{span}\{S_2^{(j)}, \widehat{S}_2^{(j)}\}$ is

$$\mathcal{S}_2^{(j)} = \operatorname{span}\{s_{n_j/2}^{(2,j)}, \ldots, s_{n_j}^{(2,j)}, \hat{s}_1^{(2,j)}, \ldots, \hat{s}_{n_j/2}^{(2,j)}\}.$$

Let $\mathcal{S}_3 \equiv \operatorname{span}\{S_3\}$ denote a maximal M-stable isotropic subspace of $\operatorname{span}\{S_3, \widehat{S}_3\}$. Since $\mathcal{S}_1^{(j)}$, $j = 1, \ldots, k_1$, $\mathcal{S}_2^{(j)}$, $j = 1, \ldots, k_2$, and $\mathcal{S}_3$ are uniquely determined with lowest Jordan degree by collecting these M-stable isotropic subspaces and letting

$$\mathcal{Y}_{\mathcal{L}} = \left(\bigoplus_{j=1}^{k_1} \mathcal{S}_1^{(j)}\right) \oplus \left(\bigoplus_{j=1}^{k_2} \mathcal{S}_2^{(j)}\right) \oplus \mathcal{S}_3,$$

we get that $\mathcal{Y}_{\mathcal{L}}$ is the M-stable Lagrangian subspace satisfying (R1).

From the above discussion, we see that the desired Lagrangian subspace $\mathcal{Y}_{\mathcal{L}}$ is spanned by the Jordan vectors corresponding to each first half of Jordan blocks of purely imaginary eigenvalue and the Jordan vectors corresponding to eigenvalues with negative real parts.

Assumption (A1) is necessary for the uniqueness of (R1). If we relax (A1) in that some partial multiplicities of zero eigenvalues of M are permitted to be odd, then the M-stable Lagrangian subspace still exists, but the uniqueness of (R1) does not hold. For example, let $M = \mathrm{diag}\{J^{(3)}(0), -J^{(3)}(0)^T\}$. Then M has zero eigenvalue with partial multiplicities 3, 3. It is easily seen that $\{e_1, e_6, e_5\}$, $\{e_1, e_2, e_6\}$, $\{e_6, e_5, e_4\}$, and $\{e_1, e_2, e_3\}$ are four distinct M-stable Lagrangian subspaces, but the first two have the same lowest Jordan degrees. As mentioned in section 1, if some nonzero eigenvalue has odd partial multiplicities, then the existence of M-stable Lagrangian subspace can fail. Let $M = \left[\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right]$. Then M has eigenvalues $\pm i$ associated with eigenvectors $\left[\begin{smallmatrix} 1 \\ i \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 1 \\ -i \end{smallmatrix}\right]$, respectively. It is easy to verify that M has no M-stable Lagrangian subspace.

The following theorem of [14] states an important result from linear-quadratic control problems.

THEOREM 2.2. *Let M be a Hamiltonian matrix as in* (1.1). *Let G be positive semidefinite and (A,G) be stabilizable. Assume* (A1) *holds. Then there exists a symplectic matrix S such that $\Lambda^{(m_j)}(0)$ in* (2.2) *has zeros everywhere except one in the (m_j, m_j)th entry. Furthermore,*

(i) *there exists a unique M-stable Lagrangian subspace $\mathcal{Y}_{\mathcal{L}}$,*

(ii) *there exists a unique symmetric solution $X \in \mathbf{R}^{n\times n}$ of CARE in* (1.4) *such that* $\mathrm{Re}(\lambda(A + GX)) \le 0$ *and* $\mathrm{span}\{\left[\begin{smallmatrix} I \\ X \end{smallmatrix}\right]\} = \mathcal{Y}_{\mathcal{L}}$. □

Remark. For the case of Theorem 2.2, the only possible M-stable isotropic subspace corresponding to zero eigenvalues must have the form (2.3). The M-stable Lagrangian subspace $\mathcal{Y}_{\mathcal{L}}$ is then uniquely determined. Thus, requirement (R1) for $\mathcal{Y}_{\mathcal{L}}$ here is automatically satisfied.

For the symplectic pencil $N - \lambda L$, we want to find the (N, L)-stable Lagrangian subspace $\mathcal{W}_{\mathcal{L}}$. By a skillful transformation of [20, p. 120], we can deflate zero and infinity eigenvalues of $N - \lambda L$ simultaneously and obtain a reduced symplectic pencil $\widehat{N} - \lambda\widehat{L}$ having only nonzero finite eigenvalues. Thus, computing the (N, L)-stable Lagrangian subspace is equivalent to computing the stable Lagrangian subspace of the symplectic matrix $B = \widehat{L}^{-1}\widehat{N}$. It is easily seen that the Cayley transformation matrix

$$M = (I + B)(I - B)^{-1} \tag{2.5}$$

is Hamiltonian. Furthermore, since the transformation (2.5) is rational and M, B are commuted, an M-stable Lagrangian subspace must be a stable Lagrangian subspace of B. Similar to the continuous-time case, we can conclude that the (N, L)-stable Lagrangian subspace $\mathcal{W}_{\mathcal{L}}$ is unique determined if (A2) and (R2) are satisfied.

Hereafter, for brevity, M-stable and (N, L)-stable Lagrangian subspaces mean the M-stable and the (N, L)-stable Lagrangian subspaces with lowest Jordan degrees, respectively.

3. Computing the stable Lagrangian subspace of a Hamiltonian matrix having purely imaginary eigenvalues. Let M be the Hamiltonian matrix as in (1.1). Assume (A1) holds; i.e., the partial multiplicities of purely imaginary eigenvalues of M are all even. In this section, we shall develop a reliable algorithm to compute

the M-stable isotropic subspace $\mathcal{Y}$ corresponding to each first half of Jordan blocks of all purely imaginary eigenvalues and get a reduced Hamiltonian matrix having no purely imaginary eigenvalue. Combining $\mathcal{Y}$ with the isotropic subspace corresponding to the strictly stable eigenvalues of M, we obtain the desired M-stable Lagrangian subspace $\mathcal{Y}_{\mathcal{L}}$.

The main idea of our algorithm to determine $\mathcal{Y}$ is that we first compute a Jordan basis corresponding to nonpositive eigenvalues of M^2 and then use it to determine a Jordan basis corresponding to purely imaginary eigenvalues of M and to determine an isotropic basis Υ of $\mathcal{Y}$.

We now consider the case of nonzero purely imaginary eigenvalues. Assume that the conjugate eigenvalue pair $\pm i\omega$ of M have the Jordan blocks $\{J^{(2m_1)}(i\omega), \ldots, J^{(2m_k)}(i\omega)\}$ and $\{J^{(2m_1)}(-i\omega), \ldots, J^{(2m_k)}(-i\omega)\}$ with even orders, respectively. It is easily seen that the negative eigenvalue $-\omega^2$ of M^2 has the Jordan blocks $\{J^{(2m_j)}(-\omega^2), J^{(2m_j)}(-\omega^2)\}_{j=1}^{k}$. Hence, the eigenspace of M corresponding to each first half of Jordan blocks $\{J^{(m_j)}(\pm i\omega)\}_{j=1}^{k}$ is just the eigenspace of M^2 corresponding to each first half of Jordan blocks $\{J^{(m_j)}(-\omega^2), J^{(m_j)}(-\omega^2)\}_{j=1}^{k}$. Thus, the desired M-stable isotropic subspace can be determined directly from the associated eigenspace of M^2. The case of zero purely imaginary eigenvalue is more complicated than the case of nonzero purely imaginary eigenvalue. In the following we shall discuss this case carefully.

Let $\{2m_1, \ldots, 2m_k\}$ with $m_1 \le \cdots \le m_k$ be the partial multiplicities of the zero eigenvalues of M and $n_0 = 2\sum_{j=1}^{k} m_j$ be the algebraic multiplicity of zero eigenvalues. Let

$$Y = \left[Y_1^{(0)}, \ldots, Y_{2m_k}^{(k-1)}\right], \tag{3.1a}$$

be an orthonormal basis of the subspace spanned by the associated Jordan vectors, where the submatrix $Y_p^{(j)}$ for $p = 1, \ldots, 2m_k$ is a $2n \times (k-j)$ orthonormal matrix of degree p and $j \equiv j(p) \in \{0, \ldots, k-1\}$ is an integer function in p such that

$$2m_j < p \le 2m_{j+1} \ (m_0 = 0). \tag{3.1b}$$

Remark. (i) A matrix Y_p is of degree p if any nonzero vector $v \in \operatorname{span}\{Y_p\}$ satisfies $M^p v = 0$ and $M^{p-1} v \neq 0$. (ii) Since the mutually orthogonal subspaces spanned by $\{Y_p^{(j)}\}$ are unique ($p = 1, \ldots, 2m_k$), for convenience we identify any two orthonormal bases of $\operatorname{span}\{Y_p^{(j)}\}$.

Furthermore, we define

$$\widetilde{Y}_p^{(j)} = \left[Y_1^{(0)}, \ldots, Y_p^{(j)}\right] \tag{3.2}$$

as the submatrix of Y of degree less than or equal to p. From elementary algebra theory, we see that the partial multiplicities of zero eigenvalues of M^2 are $\{m_1, m_1, \ldots, m_k, m_k\}$. Let

$$Z = \left[Z_1^{(0)}, \ldots, Z_{m_k}^{(k-1)}\right] \tag{3.3a}$$

be an orthonormal basis of the associated Jordan vectors, where the submatrix $Z_p^{(j)}$ for $p = 1, \ldots, m_k$ is a $2n \times 2(k-j)$ orthonormal matrix of degree p and $j \equiv j(p) \in \{0, \ldots, k-1\}$ is an integer function in p such that

$$m_j < p \le m_{j+1} \ (m_0 = 0). \tag{3.3b}$$

We also define

$$\widetilde{Z}_p^{(j)} = \left[Z_1^{(0)}, \ldots, Z_p^{(j)} \right] \tag{3.4}$$

as the submatrix of Z of degree less than or equal to p. Let Υ_s be an orthonormal isotropic subbasis corresponding to each first half of Jordan blocks of zero eigenvalues. In fact, Υ_s here is an orthonormal basis of the maximal isotropic subspace corresponding to zero eigenvalues and $\text{span}\{\Upsilon_s\} \subset \mathcal{Y}_{\mathcal{L}}$. The approach of our algorithm is that we use Z to determine Y and then use Y to compute Υ_s.

We now develop a reliable algorithm to compute the matrix Z described in (3.3a,b). For convenience hereafter, we assume that the only purely imaginary eigenvalue of M is zero.

ALGORITHM 3.1. *This algorithm computes an orthonormal subbasis* $Z = [Z_1^{(0)}, \ldots, Z_{m_k}^{(k-1)}]$ *of* M^2 *corresponding to zero eigenvalues.*

Step 1: *Reduce* M^2 *to a Hessenberg matrix by using the squared reduced algorithm of* [27]. *That is, find a* $2n \times 2n$ *symplectic orthogonal matrix* Q *so that*

$$Q^T M^2 Q = H \equiv \begin{bmatrix} H_1 & K_1 \\ O & H_1^T \end{bmatrix},$$

where H_1 *is upper Hessenberg and* K_1 *is skew-symmetric.*

Step 2: *Reduce* H_1 *to a real Schur form by the QR algorithm, e.g.,* [9, *p.* 228]. *That is, find an* $n \times n$ *orthogonal matrix* Q_1 *so that*

$$Q_1^T H_1 Q_1 = R_1, \qquad Q_1^T K_1 Q_1 = S_1,$$

where R_1 *is quasi-upper triangular.*
Let $n_0 =$ *the algebraic multiplicity of zero eigenvalues of* M^2.
Let

$$H := \begin{bmatrix} I & O \\ O & \hat{I} \end{bmatrix} \begin{bmatrix} R_1 & S_1 \\ O & R_1^T \end{bmatrix} \begin{bmatrix} I & O \\ O & \hat{I} \end{bmatrix} \quad \textit{(quasi-upper triangular),}$$

$$Q := \begin{bmatrix} Q_1 & O \\ O & Q_1 \end{bmatrix} \begin{bmatrix} I & O \\ O & \hat{I} \end{bmatrix}, \quad \textit{where } \hat{I} = \begin{bmatrix} 0 & & 1 \\ & \cdot & \\ 1 & & 0 \end{bmatrix}.$$

Set $E := I_{2n}$, $j = 0$, $q = 1$, *and* $m_0 = 0$.

Step 3: Repeat:

3.1 *Find an orthonormal basis* $\widehat{B}_0$ *of null space of* H *by applying an RRQR factorization of* [4]. *That is, find a permutation* Π_1 *and an orthogonal matrix* V_1 *such that*

$$\Pi_1 H V_1 = \begin{pmatrix} O & X \\ O & \widehat{H} \end{pmatrix},$$

where $\widehat{H}$ *is quasi-upper triangular. Let* γ_H *be the nullity of* H.
Set $\widehat{B}_0 = V_1 \begin{bmatrix} I_{\gamma_H} \\ 0 \end{bmatrix}$.
Comment: *An RRQR factorization of a quasi-upper triangular* H *needs only* $O(n^2)$ *flops if* $n_0 \ll n$.

- *If* $q = 1$, *then*

$$k = \frac{\gamma_H}{2}, \quad \gamma^* = \gamma_H, \quad Jump = 0, \quad B_0 = \widehat{B}_0,$$

else

$$\gamma = \gamma_H, \quad Jump = \gamma^* - \gamma, \quad B_0 = \begin{bmatrix} 0 \\ \widehat{B}_0 \end{bmatrix} \in \mathbf{R}^{2n \times \gamma_H}$$

- *If* $Jump \neq 0$, *then for* $\ell = j+1, \ldots, j + \frac{Jump}{2}$, *set* $m_\ell = q - 1$ *and update* $j = j + \frac{Jump}{2}$, $\gamma^* = \gamma$.
- *Set* $Z_q^{(j)} = QB_0$.
- *If* $j = k$, *then stop.*

3.2 *Find two orthogonal matrices* U_2 *and* V_2 *by using Algorithm* 3.1.1 *proposed by Beelen and Van Dooren* [3] *such that*

$$U_2^T \left(\Pi_1 H V_1\right) V_2 = \left[\begin{array}{c|c} 0 & H_{12} \\ \hline 0 & H_{22} \end{array}\right], \quad U_2^T \left(\Pi_1 E V_1\right) V_2 = \left[\begin{array}{c|c} E_{11} & E_{12} \\ \hline 0 & E_{22} \end{array}\right].$$

Comment: (i) *Here the matrix* H_{22} *is preserved to be quasi-upper triangular and* E_{11} *is nonsingular. Algorithm* 3.1.1 *of* [3] *used in Step* 3.2 *needs only* $O(n^2)$ *flops.* (ii) *This substep determines the partial multiplicities and an orthonormal basis for the associated Jordan vectors* [3].

3.3 *Update (deflation step):*

- $H := H_{22}$ *(dimension reduced).*
- $E := E_{22}$ *(dimension reduced).*
- *If* $q = 1$, *then set* $Q = Q(V_1 V_2)$,
  else set $Q = Q \left[\begin{array}{c|c} I & 0 \\ \hline 0 & V_1 V_2 \end{array}\right] \in \mathbf{R}^{2n \times 2n}$.
- *Set* $q = q + 1$, *go to* **Repeat**. □

Remark. (i) Instead of Step 3.2, one can also use a nonequivalence transformation to deflate the zero eigenvalues of the pencil $H - \lambda E$ [17]. The algorithm uses nonunitary transformations but needs only about one-fourth flops of Algorithm 3.1.1 of [3]. (ii) If M^2 has a negative eigenvalue $-\omega^2$, then we replace the matrix H in Step 3 by $H + \omega^2 I$ and perform the same process to compute an associated Jordan basis corresponding to $-\omega^2$. (iii) This algorithm uses only orthogonal transformations. The accuracy of the computed orthonormal Jordan subbasis Z depends on the sensitivity of the computed nonpositive eigenvalues $-\omega^2$ of M^2. It is shown in [27] that the computed $\pm i\omega$ are the exact eigenvalues of a matrix $M + E$ where $\|E\|$ depends on the square root of the machine precision. Hence, the accuracy of the computed Z is reliable when the sensitivity of $\pm i\omega$ of M is acceptable.

The following theorem gives the relation between orthonormal Jordan bases corresponding to zero eigenvalues of M^2 and M, respectively. We use the notation defined in (3.1)–(3.4) but omit the superscript (j).

Let $\widetilde{Z} = [Z_1, \ldots, Z_q]$ and $\widetilde{Y}_p = [Y_1, \ldots, Y_p]$, where Z_q and Y_p are orthonormal Jordan bases of M^2 and M, respectively, of degree q and p for $q = 1, \ldots, m_k$ and $p = 1, \ldots, 2m_k$.

THEOREM 3.2. *For* $p = 1, \ldots, m_k$, *it holds that*

(i) $\operatorname{span}\{\widetilde{Y}_{2p}\} = \operatorname{span}\{\widetilde{Z}_p\}$,

(ii) $\operatorname{span}\{Z_p\} = \operatorname{span}\{Y_{2p-1}\} \oplus \operatorname{span}\{Y_{2p}\}$,

(iii) $\text{span}\{Y_{2p-1}\} = \text{span}\{(I - \widetilde{Z}_{p-1}\widetilde{Z}_{p-1}^T)MZ_p\} = \text{span}\{(Z_pZ_p^T)MZ_p\}$,

(iv) *if* W_{2p-1} *is an orthonormal basis of* $\text{span}\{(I - \widetilde{Z}_{p-1}\widetilde{Z}_{p-1}^T)MZ_p\}$, *then* $\text{span}\{Y_{2p}\} = \text{span}\{(I - W_{2p-1}W_{2p-1}^T)Z_p\}$.

For convenience, here we use $\widetilde{Z}_0 = 0$.

Proof. (i) Since $(M^2)^p v = M^{2p}v$ for any $v \in \mathbf{R}^{2n\times 1}$, (i) follows.

(ii) From (i), we have

$$\text{span}\{Y_{2p}\} \oplus \text{span}\{Y_{2p-1}\} \oplus \text{span}\left\{\widetilde{Y}_{2p-2}\right\} = \text{span}\{Z_p\} \oplus \text{span}\left\{\widetilde{Z}_{p-1}\right\}.$$

Furthermore, both subspaces $\text{span}\{Z_p\}$ and $\text{span}\{Y_{2p}\} \oplus \text{span}\{Y_{2p-1}\}$ are orthogonal to $\text{span}\{\widetilde{Y}_{2p-2}\}$ (i.e., $\text{span}\{\widetilde{Z}_{p-1}\}$). Hence, (ii) is proved.

(iii) By the definition of Z_p, we have

$$\text{span}\{MZ_p\} \subset \text{span}\left\{\widetilde{Y}_{2p-1}\right\} = \text{span}\{Y_{2p-1}\} \oplus \text{span}\left\{\widetilde{Z}_{p-1}\right\}. \tag{3.5}$$

This implies that

$$\text{span}\left\{\left(I - \widetilde{Z}_{p-1}\widetilde{Z}_{p-1}^T\right)MZ_p\right\} \subset \text{span}\{Y_{2p-1}\}. \tag{3.6}$$

On the other hand, from (ii), we have

$$\text{span}\left\{\left(I - \widetilde{Z}_{p-1}\widetilde{Z}_{p-1}^T\right)MY_{2p}\right\} \subset \text{span}\left\{\left(I - \widetilde{Z}_{p-1}\widetilde{Z}_{p-1}^T\right)MZ_p\right\}. \tag{3.7}$$

By (3.6) and (3.7), it is easily seen that

$$\begin{aligned}\dim\left(\text{span}\left\{\left(I - \widetilde{Z}_{p-1}\widetilde{Z}_{p-1}^T\right)MY_{2p}\right\}\right) &\le \dim\left(\text{span}\left\{\left(I - \widetilde{Z}_{p-1}\widetilde{Z}_{p-1}^T\right)MZ_p\right\}\right)\\ &\le \dim\left(\text{span}\{Y_{2p-1}\}\right).\end{aligned} \tag{3.8}$$

From (3.6) and (3.7), it follows that to verify the first equality of (iii) it is sufficient to show that both inequalities in (3.8) hold. Now, suppose that

$$\dim\left(\text{span}\left\{\left(I - \widetilde{Z}_{p-1}\widetilde{Z}_{p-1}^T\right)MY_{2p}\right\}\right) < \dim\left(\text{span}\{Y_{2p-1}\}\right). \tag{3.9}$$

Since all partial multiplicities of zero eigenvalues are even,

$$\dim\left(\text{span}\{Y_{2p}\}\right) = \dim\left(\text{span}\{Y_{2p-1}\}\right). \tag{3.10}$$

From (3.9) and (3.10) it follows that the column vectors of $(I - \widetilde{Z}_{p-1}\widetilde{Z}_{p-1}^T)MY_{2p}$ are linearly dependent. Thus, there exists a nonzero vector ξ such that

$$\left(I - \widetilde{Z}_{p-1}\widetilde{Z}_{p-1}^T\right)MY_{2p}\xi = 0.$$

This implies that $MY_{2p}\xi \in \text{span}\{\widetilde{Z}_{p-1}\}$. By the definition of $\widetilde{Z}_{p-1}$, we then have

$$M^{2p-1}Y_{2p}\xi = (M^2)^{p-1}MY_{2p}\xi = 0. \tag{3.11}$$

This contradicts the definition of Y_{2p}. Therefore, the strict inequality in (3.9) does not hold; i.e., both equalities in (3.8) hold. Thus, the first equality of (iii) is proved.

From (ii), we know that there exists an orthonormal matrix U such that

$$Z_p = [Y_{2p-1}, Y_{2p}]\, U. \tag{3.12}$$

This implies that

$$MZ_p = [MY_{2p-1}, MY_{2p}]\, U. \tag{3.13}$$

On the other hand, by the definitions of Y_{2p-1} and $\widetilde{Z}_{p-1}$, we have

$$\operatorname{span}\{MY_{2p-1}\} \subset \operatorname{span}\left\{\widetilde{Z}_{p-1}\right\}. \tag{3.14}$$

From (3.13) and (3.14),

$$\left(Z_p Z_p^T\right) MZ_p = \left[0,\, Z_p Z_p^T MY_{2p}\right] U.$$

Hence, we get

$$\operatorname{span}\left\{\left(Z_p Z_p^T\right) MZ_p\right\} = \operatorname{span}\left\{\left(Z_p Z_p^T\right) MY_{2p}\right\}.$$

Furthermore, from (3.5) and (3.12), we have

$$\operatorname{span}\left\{\left(Z_p Z_p^T\right) MY_{2p}\right\} = \operatorname{span}\left\{\left(Z_p Z_p^T\right) MZ_p\right\} \subset \operatorname{span}\{Y_{2p-1}\}.$$

This implies

$$\dim\left(\operatorname{span}\left\{\left(Z_p Z_p^T\right) MY_{2p}\right\}\right) \le \dim\left(\operatorname{span}\{Y_{2p-1}\}\right). \tag{3.15}$$

Suppose the inequality of (3.15) holds. Then, from (3.10), we conclude that there exists a vector $\xi \neq 0$ such that

$$\left(Z_p Z_p^T\right) MY_{2p}\xi = 0.$$

This implies $MY_{2p}\xi \in \operatorname{span}\{\widetilde{Z}_{p-1}\}$. By the same argument as (3.11) we get the contradiction. Therefore, the second equality of (iii) is proved.

(iv) From (ii) and (iii) immediately follows (iv). □

Remark. From statements (iii) and (iv) of Theorem 3.2, we see that the matrices Y_{2p-1} and Y_{2p} can be replaced by an orthonormal basis of $\operatorname{span}\{(Z_p Z_p^T)MZ_p\}$ and $\operatorname{span}\{(I - W_{2p-1}W_{2p-1}^T)Z_p\}$, respectively. In the following, we develop an algorithm for computing Y_{2p-1} and Y_{2p} by using the orthonormal bases Z_p.

ALGORITHM 3.3. *This algorithm computes Y_{2p-1} and Y_{2p} by using the orthonormal basis Z_p, $p = 1, \ldots, m_k$, obtained by Algorithm* 3.1.

Step 1. *Compute an orthonormal basis $Q_1^{(0)}$ of $MZ_1^{(0)}$ and set*

$$Y_1^{(0)} = Q_1^{(0)}.$$

Step 2. *Compute the SVD of $(Q_1^{(0)})^T Z_1^{(0)}$ such that*

$$\left(U_1^{(0)}\right)^T \left(\left(Q_1^{(0)}\right)^T Z_1^{(0)}\right) V_1^{(0)} = \left[\begin{array}{c|c} \Sigma_1^{(0)} & 0 \end{array}\right],$$

where $U_1^{(0)}, V_1^{(0)}$ *are two unitary matrices and*

$$\Sigma_1^{(0)} = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}.$$

Set

$$Y_2^{(0)} = Z_1^{(0)} V_1^{(0)} \begin{bmatrix} 0 \\ \hline I_k \end{bmatrix}.$$

Set $p = 2$.

Step 3. Repeat:

If $p > \frac{m_k}{2} + 1$, *then stop.*

Determine $j \in \{0, 1, \ldots, k-1\}$ *such that* $m_j < p \le m_{j+1}$.

3.1 *Compute an orthonormal basis* $Q_p^{(j)}$ *of* $MZ_p^{(j)}$.

3.2 *Compute the SVD of* $(Z_p^{(j)})^T Q_p^{(j)}$ *such that*

$$\left(U_{2p-1}^{(j)}\right)^T \left[\left(Z_p^{(j)}\right)^T Q_p^{(j)}\right] V_{2p-1}^{(j)} = \Sigma_{2p-1}^{(j)},$$

where $U_{2p-1}^{(j)}$, $V_{2p-1}^{(j)}$ *are two unitary matrices and*

$$\Sigma_{2p-1}^{(j)} = \left[\begin{array}{ccc|c} \sigma_1^{(2p-1)} & & & \\ & \ddots & & \\ & & \sigma_{k-j}^{(2p-1)} & \\ \hline & & & O_{k-j} \end{array}\right]$$

with $\sigma_1^{(2p-1)} \ge \cdots \ge \sigma_{k-j}^{(2p-1)} > 0$.

Set

$$Y_{2p-1}^{(j)} = Z_p^{(j)} U_p^{(j)} \begin{bmatrix} I_{k-j} \\ \hline 0 \end{bmatrix}.$$

3.3 *Compute the SVD of* $\left(Y_{2p-1}^{(j)}\right)^T Z_p^{(j)}$ *such that*

$$\left(U_{2p}^{(j)^T}\right) \left[\left(Y_{2p-1}^{(j)}\right)^T Z_p^{(j)}\right] V_{2p}^{(j)} = \left[\Sigma_{2p}^{(j)} \middle| O\right],$$

where $U_{2p}^{(j)}$, $V_{2p}^{(j)}$ *are two unitary matrices and*

$$\Sigma_{2p}^{(j)} = \begin{bmatrix} \sigma_1^{(2p)} & & \\ & \ddots & \\ & & \sigma_{k-j}^{(2p)} \end{bmatrix}.$$

Set

$$Y_{2p}^{(j)} = Z_p^{(j)} V_{2p}^{(j)} \begin{bmatrix} 0 \\ \hline I_{k-j} \end{bmatrix}.$$

3.4 *Update* $p := p + 1$ *and go to* **Repeat**.

This algorithm needs about $O(n^2)$ *flops.* □

Denote Υ_s as an orthonormal basis of the M-stable isotropic subspace corresponding to the first half of Jordan blocks of zero eigenvalues. We now define a sequence of orthonormal bases $\{\widetilde{\Upsilon}_p\}_{p=1}^{m_k}$ which is closely related to the matrix Υ_s.

DEFINITION 3.4. *Let* $\widetilde{\Upsilon}_p$ *for* $p = 1, \ldots, m_k$ *be a maximal orthonormal basis satisfying the following:*

(i) $\text{span}\{\widetilde{\Upsilon}_p\} \subset \mathcal{N}(M^p)$ *(null space of* M^p*).*

(ii) $x^T J y = 0$ *for any* x, $y \in \text{span}\{\widetilde{\Upsilon}_p\}$.

(iii) $\text{span}\{\widetilde{\Upsilon}_{p-1}\} \subset \text{span}\{\widetilde{\Upsilon}_p\}$. *(Here,* $\widetilde{\Upsilon}_0 \equiv 0$*.)*

(iv) *If there is a subspace* $\mathcal{V} \subset \mathbf{R}^{2n}$ *satisfying statements* (i), (ii), *and* (iii), *then* $\mathcal{V} \subset \text{span}\{\widetilde{\Upsilon}_p\}$.

THEOREM 3.5. *The following properties for the sequence* $\{\widetilde{\Upsilon}_p\}_{p=1}^{m_k}$ *defined above are true:*

(i) $\text{span}\{\widetilde{\Upsilon}_p\}$ *is unique for* $p \in \{1, \ldots, m_k\}$.

(ii) $\text{span}\{\widetilde{\Upsilon}_{m_k}\} = \text{span}\{\Upsilon_s\}$.

Proof. (i) From Theorem 2.1 and assumption (A1), we can assume that M has the form (2.2). Since $\text{span}\{\widetilde{\Upsilon}_p\} \subset \mathcal{N}(M^p)$ for $p = 1, \ldots, m_k$, for convenience, we assume without loss of generality (w.l.o.g.) that M has only zero eigenvalues and discuss two typical cases of M in the following.

Case 1. Let $k = 2$, $m_1 < m_2$, and

$$M = \left[\begin{array}{cc|cc} J^{(m_1)}(0) & & \Lambda^{(m_1)}(0) & \\ & J^{(m_2)}(0) & & \Lambda^{(m_2)}(0) \\ \hline & & -J^{(m_1)}(0)^T & \\ & & & -J^{(m_2)}(0)^T \end{array}\right]$$

with $\Lambda^{(m_1)}(0)(m_1, m_1) = \Lambda^{(m_2)}(0)(m_2, m_2) = 1$.

For $p \le m_1$, we have

$$\mathcal{N}(M^p) = \text{span}\{e_1, \ldots, e_p, e_{m_1+1}, \ldots, e_{m_1+p}\}.$$

Since $p < m_1 + p \le m_1 + m_2$ (= the half of dimension of M) for any x, $y \in \mathcal{N}(M^p)$ we have $x^T J y = 0$. From the definition of $\widetilde{\Upsilon}_p$ it follows that

$$\text{span}\{\widetilde{\Upsilon}_p\} = \mathcal{N}(M^p).$$

In addition, $\mathcal{N}(M^p)$ is unique. Thus, $\text{span}\{\widetilde{\Upsilon}_p\}$ is unique for $p \le m_1$.

For $m_1 + 1 \le p \le m_2$, we have

$$\mathcal{N}(M^p) = \text{span}\{e_1, \ldots, e_{m_1}, e_{(m_1+m_2)+m_1}, \ldots, e_{(m_1+m_2)+m_1-p+1}, e_{m_1+1}, \ldots, e_{m_1+p}\}.$$

Let $\mathcal{U} \equiv \text{span}\{e_1, \ldots, e_{m_1}, e_{m_1+1}, \ldots, e_{m_1+p}\}$. Obviously, $\mathcal{U} \subset \mathcal{N}(M^p)$. Since $m_1 + p < m_1 + m_2$ for any x, $y \in \mathcal{U}$ we have $x^T J y = 0$. Hence,

$$\mathcal{U} \subset \text{span}\{\widetilde{\Upsilon}_p\}.$$

If $\mathcal{U} \neq \text{span}\{\widetilde{\Upsilon}_p\}$, then there exists a nonzero vector $v \in \text{span}\{e_{(m_1+m_2)+m_1}, \ldots, e_{(m_1+m_2)+m_1-p+1}\}$ such that $v \in \text{span}\{\widetilde{\Upsilon}_p\}$ and $v \notin \mathcal{U}$. But, for this v, there exists an associated nonzero vector $u \in \mathcal{U}$ such that

$$u^T J v \neq 0.$$

This contradicts the definition of $\mathrm{span}\{\widetilde{\Upsilon}_p\}$. Hence $\mathcal{U} = \mathrm{span}\{\widetilde{\Upsilon}_p\}$. Since $\mathcal{U}$ is unique, the proof follows.

Case 2. Let $k = 3$, $2m_1 < m_2$, and

$$M = \left[\begin{array}{cc|cc} J^{(2m_1)}(0) & & 0 & \\ & J^{(m_2)}(0) & & \Lambda^{(m_2)}(0) \\ \hline & & -J^{(2m_1)}(0)^T & \\ & & & -J^{(m_2)}(0)^T \end{array}\right]$$

with $\Lambda^{(m_2)}(0)(m_2, m_2) = 1$. The proof of this case is similar to that for Case 1. We omit it here.

(ii) By the definition of $\mathrm{span}\{\widetilde{\Upsilon}_p\}$ and (i), (ii) follows immediately. □

Remark. If we ignore the monotone property of $\mathrm{span}\{\widetilde{\Upsilon}_p\}$, i.e., condition (iii) of Definition 3.4, then the uniqueness of $\mathrm{span}\{\widetilde{\Upsilon}_p\}$ does not hold. For example, let $2m_1 = 2$, $2m_2 = 4$, and

$$M = \left[\begin{array}{c|ccc|cc} 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{array}\right];$$

then for $p = 2$ there exist two different maximal isotropic orthonormal bases $\{e_1^{(6)}, e_2^{(6)}, e_3^{(6)}\}$ and $\{e_2^{(6)}, e_3^{(6)}, e_4^{(6)}\}$. But the latter does not form a subspace of $\mathrm{span}\{\Upsilon_s\}$. Hence, we must determine Υ_s by using a monotone process.

We now develop an algorithm to determine the maximal isotropic subbasis Υ_s by using the computed $\widetilde{Y}_{m_k} \equiv [Y_1, \ldots, Y_{m_k}]$ and Theorem 3.5.

ALGORITHM 3.6. *This algorithm computes Υ_s by using orthonormal basis $\widetilde{Y}_{m_k}$ obtained by Algorithm* 3.3.

Step 1. *Let $\widehat{\Upsilon} = [Y_1^{(0)}, \ldots, Y_{m_1}^{(0)}]$ and $\widehat{p} = m_1 + 1$.*

Step 2. Repeat:

Determine $j \in \{1, \ldots, k\}$ is a maximal integer such that $m_j < \widehat{p}$.

If $j = k$, set $\Upsilon_s = \widehat{\Upsilon}$ and stop.

For $p = m_j + 1, \ldots, m_{j+1}$:

2.1 *Find $i \geq 0$ such that $2m_i < p \leq 2m_{i+1}$.*

If $p = m_1 + 1$, then $\widehat{Y}_{m_1+1}^{(0)} = Y_{m_1+1}^{(0)}$,

else $\widehat{Y}_p^{(i)} = [Y_{m_1+1}^{(0)}, \ldots, Y_p^{(i)}]$.

Let #1 = the number of columns of $\widehat{\Upsilon}$.

Let #2 = the number of columns of $\widehat{Y}_p^{(i)}$.

2.2 *Compute the SVD of $\widehat{\Upsilon}^T J \widehat{Y}_p^{(i)}$ such that*

$$\left(U_p^{(j)}\right)^T \left[\widehat{\Upsilon}^T J \widehat{Y}_p^{(i)}\right] V_p^{(j)} = \left[\Sigma_p^{(j,i)} \middle| O\right],$$

where

$$\Sigma_p^{(j,i)} = \left[\begin{array}{ccc} \sigma_1^{(j,i)} & & \\ & \ddots & \\ & & \sigma_{\#1}^{(j,i)} \end{array}\right].$$

Let $\#3 = \max\{q|\sigma_q^{(j,i)} > 0 \text{ for } q = 1, \ldots, \#1\}$.

2.3 *Update* $\widehat{\Upsilon} = [Y_1^{(0)}, \ldots, Y_{m_1}^{(0)}, \widehat{Y}_p^{(i)} V_p^{(j)} [\frac{0}{I_{\#2-\#3}}]]$.

Endfor.

Step 3. *Update* $\widehat{p} = p + 1$ *and go to* **Repeat**.

Comment: (i) *In substep* 2.1*, it is easily seen that*

$$\#1 = \sum_{\ell=0}^{j-1} (m_{\ell+1} - m_\ell)(k - \ell) + (p - m_j)(k - j)$$

and

$$\#2 = 2\sum_{\ell=0}^{i-1} (m_{\ell+1} - m_\ell)(k - \ell) + (p - 2m_i)(k - i) - m_1 k.$$

(ii) *This algorithm needs about* $O(n^2)$ *flops.* □

After the M-stable isotropic subspace $\text{span}\{\Upsilon_s\}$ is found, we can deflate it by using symplectic orthogonal transformations to get a reduced Hamiltonian matrix $\widehat{M}$ (say!) having no purely imaginary eigenvalue. Then we compute the maximal stable isotropic subspace of $\widehat{M}$ by exploiting [2, 23, 29]. Combining these two computed isotropic subspaces, we obtain the desired M-stable Lagrangian subspace $\mathcal{Y}_\mathcal{L}$.

4. Computing the stable Lagrangian subspace of a symplectic pencil having unimodular eigenvalues. Let $N - \lambda L$ be a symplectic pencil as in (1.2). Assume (A2) holds; i.e., the partial multiplicities of unimodular eigenvalues of $N - \lambda L$ are all even. In this section, we shall develop an algorithm to compute the (N, L)-stable isotropic subspace $\mathcal{W}$ corresponding to the first half Jordan blocks of all unimodular eigenvalues and get a reduced symplectic pencil having no unimodular eigenvalue. Combining $\mathcal{W}$ with the maximal isotropic subspace corresponding to the strictly stable eigenvalues of $N - \lambda L$, we obtain the desired (N, L)-stable Lagrangian subspace $\mathcal{W}_\mathcal{L}$.

The main idea of our algorithm to determine $\mathcal{W}$ is that by using $S + S^{-1}$-transformation [18] we first compute a Jordan basis of $\Gamma - \lambda\Delta$ as in (1.7) corresponding to eigenvalues with magnitudes between -2 and 2 and a Jordan basis corresponding to unimodular eigenvalues of $N - \lambda L$ and then use it to determine an isotropic basis Υ of $\mathcal{W}$.

We recall from (1.7) that

$$\Gamma - \lambda\Delta \equiv \left[\left(NJL^T + LJN^T\right) - \lambda LJL^T\right] J^T.$$

Now we want to show the relation between Jordan bases corresponding to the unimodular eigenvalue μ of $N - \lambda L$ and the eigenvalue $\mu + \mu^{-1}$ of $\Gamma - \lambda\Delta$, respectively. For the pencil $N - \lambda L$, we can use the method of [20, p. 120] to deflate its zero and infinity eigenvalues simultaneously and get a reduced symplectic pencil having no zero or infinity eigenvalues. Hence, we can assume w.l.o.g. that both N and L are nonsingular in the following.

THEOREM 4.1. *Let* $N - \lambda L$ *be a symplectic pencil having unimodular eigenvalues* $\mu \in \{\pm 1, e^{\pm i\theta},\ (\theta \neq 0)\}$. *Let*

$$J^{(2m_1)}(\mu), \ldots, J^{(2m_k)}(\mu) \tag{4.1}$$

be the corresponding Jordan blocks with even sizes. Then

(i) *for* $\mu = \pm 1$ *the corresponding eigenvalue* 2 *or* -2 *of* $\Gamma - \lambda\Delta$ *has Jordan blocks*

$$J^{(m_1)}(\pm 2), J^{(m_1)}(\pm 2), \ldots, J^{(m_k)}(\pm 2), J^{(m_k)}(\pm 2),$$

(ii) *for* $\mu = e^{\pm i\theta}$ *the corresponding eigenvalue* $e^{i\theta} + e^{-i\theta}$ *of* $\Gamma - \lambda\Delta$ *has Jordan blocks*

$$J^{(2m_1)}(e^{i\theta} + e^{-i\theta}), \ldots, J^{(2m_k)}(e^{i\theta} + e^{-i\theta})$$

with the same sizes as (4.1).

Proof. To prove this theorem, we consider the following simple case. The complete proof is a straightforward generalization. Let $Y = [y_1, \ldots, y_{2m_1}]$ be a Jordan basis of $J^{(2m_1)}(\mu)$ satisfying

$$NY = LYJ^{(2m_1)}(\mu). \tag{4.2}$$

Write $Y = JL^TJ^TZ$ with $Z = [z_1, \ldots, z_{2m_1}]$. Substituting Y into (4.2), we have

$$NJL^TJ^TZ = LJL^TJ^TZJ^{(2m_1)}(\mu). \tag{4.3}$$

Since $NJN^T = LJL^T$ and N and $J^{(2m_1)}(\mu)$ are invertible, from (4.3) we get

$$LJN^TJ^TZ = LJL^TJ^TZJ^{(2m_1)}(\mu)^{-1}. \tag{4.4}$$

Combining (4.3) and (4.4), we get

$$\left(NJL^TJ^T + LJN^TJ^T\right)Z = LJL^TJ^TZ\left(J^{(2m_1)}(\mu) + J^{(2m_1)}(\mu)^{-1}\right).$$

If $\mu = \pm 1$, then it is easily seen that

$$J^{(2m_1)}(\pm 1) + J^{(2m_1)}(\pm 1)^{-1} \overset{s.}{\sim} \begin{bmatrix} \pm 2 & 0 & \pm 1 & & 0 \\ & & & \ddots & \\ & & \ddots & \ddots & \pm 1 \\ & & & & 0 \\ 0 & & & & \pm 2 \end{bmatrix}$$
$$\overset{s.}{\sim} \left[\begin{array}{c|c} J^{(m_1)}(\pm 2) & 0 \\ \hline 0 & J^{(m_1)}(\pm 2) \end{array}\right].$$

Here the symbol $\overset{s.}{\sim}$ denotes "similar." Thus, statement (i) is proved.

If $\mu = e^{\pm i\theta}$, then it is easily seen that

$$J^{(2m_1)}(e^{\pm i\theta}) + J^{(2m_1)}(e^{\pm i\theta})^{-1} \overset{s.}{\sim} \begin{bmatrix} e^{i\theta} + e^{-i\theta} & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & e^{i\theta} + e^{-i\theta} \end{bmatrix}.$$

Hence, statement (ii) follows. ∎

As in section 3, we can also give the relation between orthonormal Jordan bases corresponding to eigenvalues 2 and 1 of $\Gamma - \mu\Delta$ and $N - \lambda L$, respectively. Here we use the same notation as in section 3.

Let $\widetilde{Z}_q = [Z_1, \ldots, Z_q]$ and $\widetilde{Y}_p = [Y_1, \ldots, Y_p]$ for $q = 1, \ldots, m_k$ and $p = 1, \ldots, 2m_k$ be the orthonormal Jordan bases corresponding to 2 and 1 of $\Gamma - \mu\Delta$ and $N - \lambda L$, respectively, where Z_q and Y_p are orthonormal Jordan bases of degree q and p, respectively. We say that Z_q is of degree q if it holds

$$(\Gamma - 2\Delta)v \in \operatorname{span}\{\Delta\widetilde{Z}_{q-1}\}, \qquad (\Gamma - 2\Delta)v \notin \operatorname{span}\{\Delta\widetilde{Z}_{q-2}\} \quad (\text{for } q \geq 2)$$

for all $v \in \operatorname{span}\{Z_q\}$ and that Y_p is of degree p if it holds

$$(N - L)v \in \operatorname{span}\{L\widetilde{Y}_{p-1}\}, \qquad (N - L)v \notin \operatorname{span}\{L\widetilde{Y}_{p-2}\} \quad (\text{for } p \geq 2)$$

for all $v \in \operatorname{span}\{Y_p\}$. Here we set $\widetilde{Z}_0 = 0$ and $\widetilde{Y}_0 = 0$.

Let Θ_q be an orthonormal basis of $JL^TJ^TZ_q$ and $\widetilde{\Theta}_q = [\Theta_1, \ldots, \Theta_q]$ for $q = 1, \ldots, m_k$.

THEOREM 4.2. *For $p = 1, \ldots, m_k$, we have*

(i) $\operatorname{span}\{\widetilde{\Theta}_p\} = \operatorname{span}\{\widetilde{Y}_{2p}\}$.

(ii) $\operatorname{span}\{\Theta_p\} = \operatorname{span}\{Y_{2p-1}\} \oplus \operatorname{span}\{Y_{2p}\}$,

(iii) $\operatorname{span}\{Y_{2p-1}\} = \operatorname{span}\{(I - \widetilde{\Theta}_{p-1}\widetilde{\Theta}_{p-1}^T)(L^{-1}N - I)\Theta_p\}$
$= \operatorname{span}\{(\Theta_p{\Theta_p}^T)(L^{-1}N - I)\Theta_p\}$,

(iv) *if W_{2p-1} is an orthonormal basis of* $\operatorname{span}\{(I - \widetilde{\Theta}_{p-1}\widetilde{\Theta}_{p-1}^T)(L^{-1}N - I)\Theta_p\}$, *then* $\operatorname{span}\{Y_{2p}\} = \operatorname{span}\{(I - W_{2p-1}W_{2p-1}^T)\Theta_p\}$.

Proof. (i) Let $p = 1$. For $u \in \operatorname{span}\{\widetilde{\Theta}_1\}$ there is a vector $v \in \operatorname{span}\{\widetilde{Z}_1\}$ such that $u = JL^TJ^Tv$. Then we have $(\Gamma - 2\Delta)v = 0$. Since

$$NL^{-1}(\Gamma - 2\Delta) = (N - L)L^{-1}(N - L)JL^TJ^T \quad (\text{from (1.7)}), \tag{4.5}$$

we have

$$0 = NL^{-1}(\Gamma - 2\Delta)v = (N - L)L^{-1}(N - L)JL^TJ^Tv.$$

Hence, $u \in \operatorname{span}\{\widetilde{Y}_2\}$ and $\operatorname{span}\{\widetilde{\Theta}_1\} \subset \operatorname{span}\{\widetilde{Y}_2\}$.

Conversely, if $u \in \operatorname{span}\{\widetilde{Y}_2\}$, then

$$(N - L)L^{-1}(N - L)u = 0. \tag{4.6}$$

By (4.5) and (4.6), we have

$$NL^{-1}(\Gamma - 2\Delta)(JL^TJ^T)^{-1}u = 0.$$

Let $v = (JL^TJ^T)^{-1}u$. Since N and L are nonsingular, we have $(\Gamma - 2\Delta)v = 0$. Thus, $v \in \operatorname{span}\{\widetilde{Z}_1\}$. Statement (i) holds for $p = 1$.

Assume that statement (i) holds for $\widehat{p} = p - 1 < m_k$. For $u \in \operatorname{span}\{\widetilde{\Theta}_p\}$ there is a vector $v \in \operatorname{span}\{\widetilde{Z}_p\}$ such that $u = JL^TJ^Tv$. By (4.5) and the definition of $\widetilde{Z}_p$, we have

$$\begin{aligned}(N - L)L^{-1}(N - L)(JL^TJ^T)v &= NL^{-1}(\Gamma - 2\Delta)v \\ &= NL^{-1}\Delta(\widetilde{Z}_{p-1}\widetilde{w}_{p-1}) \\ &= N(JL^TJ^T\widetilde{Z}_{p-1}\widetilde{w}_{p-1})\end{aligned}$$

for some nonzero vector $\widetilde{w}_{p-1}$. Since (i) holds for $\widehat{p} = p - 1$, there is a nonzero vector $\widehat{w}_{2(p-1)}$ such that

$$\widetilde{Y}_{2(p-1)}\widehat{w}_{2(p-1)} = JL^TJ^T\widetilde{Z}_{p-1}\widetilde{w}_{p-1}.$$

This implies

$$\begin{aligned}(N-L)L^{-1}(N-L)(JL^TJ^T)v &= N\widetilde{Y}_{2(p-1)}\widehat{w}_{2(p-1)}\\ &= (N-L)\widetilde{Y}_{2(p-1)}\widehat{w}_{2(p-1)} + L\widetilde{Y}_{2(p-1)}\widehat{w}_{2(p-1)}\\ &\in \text{span}\{L\widetilde{Y}_{2(p-1)}\}.\end{aligned}$$

Hence, $u \in \text{span}\{\widetilde{Y}_{2p}\}$.

Conversely, if $u \in \text{span}\{\widetilde{Y}_{2p}\}$, from the proof of (i) of Theorem 4.1, we know that there is a nonzero vector v with $u = JL^TJ^Tv$ such that $v \in \text{span}\{\widetilde{Z}_p\}$. Hence, by induction, statement (i) follows.

(ii), (iii), (iv) From (i) we have that

$$\text{span}\left\{(L^{-1}N - I)\Theta_p\right\} \subset \text{span}\left\{\widetilde{Y}_{2p-1}\right\}. \tag{4.7}$$

Using (4.7) and a similar argument as in Theorem 3.2, we obtain (ii), (iii), and (iv) immediately. ∎

According to Theorems 4.1, 4.2, and 3.5, we can also develop a structure-preserving algorithm to compute the (N, L)-stable Lagrangian subspace $\mathcal{W}_{\mathcal{L}}$. The algorithm is similar to Algorithms 3.1, 3.3, and 3.6. We omit the detail descriptions while the statements are the same.

ALGORITHM 4.3. *This algorithm computes the desired (N, L)-stable isotropic basis Υ_s. Suppose that the only unimodular eigenvalue of $N - \lambda L$ is one.*

Step 1: *Reduce the skew-Hamiltonian pencil $\Gamma - \lambda\Delta \equiv [(NJL^T + LJN^T) - \lambda LJL^T]J^T$ to a skew-Hamiltonian quasi-upper upper triangular pencil by using the stable algorithm proposed by* [22]; *i.e., find orthogonal matrices U and Q such that*

$$U^T\Gamma Q = \begin{bmatrix} \Gamma_1 & H_1 \\ O & \Gamma_1^T \end{bmatrix} \equiv H$$

and

$$U^T\Delta Q = \begin{bmatrix} \Delta_1 & E_1 \\ O & \Delta_1^T \end{bmatrix} \equiv E,$$

where Γ_1 is quasi-upper triangular, Δ_1 is upper triangular, and H_1, E_1 are skew symmetric.
Set

$$H := \begin{bmatrix} I & O \\ O & \hat{I} \end{bmatrix} H \begin{bmatrix} I & O \\ O & \hat{I} \end{bmatrix} \quad \textit{(quasi-upper triangular)},$$

$$E := \begin{bmatrix} I & O \\ O & \hat{I} \end{bmatrix} E \begin{bmatrix} I & O \\ O & \hat{I} \end{bmatrix} \quad \textit{(upper triangular)},$$

$$Q := Q \begin{bmatrix} I & O \\ O & \hat{I} \end{bmatrix}, \quad \textit{where } \hat{I} = \begin{bmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{bmatrix}.$$

Let $j = 0$, $q = 1$.

Step 2: *Compute* $Z_q^{(j)}$ *for* $q = 1, \ldots, m_k$, *by performing the same statements in Step* 3 *of Algorithm* 3.1 *but replacing* H *by* $H - 2E$. *Compute an orthonormal basis* $\Theta_q^{(j)}$ *of* $JL^TJ^TZ_q^{(j)}$ *for* $q = 1, \ldots, m_k$.

Comment: *Here* $Z_q^{(j)}$ *is an orthonormal Jordan basis of degree* q *corresponding to eigenvalue* 2 *of* $\Gamma - \lambda\Delta$.

Step 3: *Perform the same statements as in Algorithm* 3.3 *but replace* $MZ_p^{(j)}$ *by* $(L^{-1}N - I)\Theta_p^{(j)}$.

Comment: *This step computes an orthonormal Jordan basis* $\{Y_p^{(j)}\}_{p=1}^{m_k}$ *of* $N - \lambda L$ *corresponding to the unimodular eigenvalue* 1.

Step 4: *Perform the same statements as in Algorithm* 3.6 *to compute the desired* (N, L)*-stable isotropic basis* Υ_s. □

Remark. If $N - \lambda L$ has unimodular eigenvalues -1 or $e^{\pm i\theta}$, then we replace $H - 2E$ in Step 2 by $H + 2E$ or $H - \eta E$ with $\eta = e^{i\theta} + e^{-i\theta}$ and perform the same process.

According to Algorithm 4.3, we can find the desired (N, L)-stable isotropic basis

$$\Upsilon_s \equiv \left[{\Upsilon_{11}^{(0)}}^T, {\Upsilon_{21}^{(0)}}^T, {\Upsilon_{31}^{(0)}}^T, {\Upsilon_{41}^{(0)}}^T \right]^T$$

with $\Upsilon_{11}^{(0)}, \Upsilon_{31}^{(0)} \in \mathbf{R}^{\frac{n_0}{2} \times \frac{n_0}{2}}$ of $N - \lambda L$. Here, n_0 is the number of unimodular eigenvalues. We now give an algorithm to determine a symplectic matrix Q and a nonsingular U such that

$$U(N - \lambda L)Q = \begin{bmatrix} N_{11} & N_{12} & 0 & 0 \\ 0 & N_{22} & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & N_{42} & 0 & I \end{bmatrix} - \lambda \begin{bmatrix} I & 0 & L_{13} & L_{14} \\ 0 & I & L_{23} & L_{24} \\ 0 & 0 & L_{33} & 0 \\ 0 & 0 & L_{43} & L_{44} \end{bmatrix}, \tag{4.8}$$

where the reduced symplectic pencil $\left[\begin{smallmatrix} N_{22} & 0 \\ N_{42} & I \end{smallmatrix}\right] - \lambda \left[\begin{smallmatrix} I & L_{24} \\ 0 & L_{44} \end{smallmatrix}\right]$ has no unimodular eigenvalue. Here $L_{44} = N_{22}^T$, $N_{42} = N_{42}^T$, and $L_{24} = L_{24}^T$.

ALGORITHM 4.4. *This algorithm is to determine a symplectic matrix* Q *and an invertible matrix* U *such that* (4.8) *holds.*

Step 1: *Find a symplectic Householder matrix* Q_1 *such that*

$$Q_1^T \Upsilon_s = \begin{bmatrix} \Upsilon_{11}^{(1)} \\ \Upsilon_{21}^{(1)} \\ \Upsilon_{31}^{(1)} \\ 0 \end{bmatrix}.$$

Set

$$N := Q_1^T N Q_1, \quad L := Q_1^T L Q_1, \quad U := Q_1^T, \quad Q := Q_1.$$

Step 2: *If* $\Upsilon_{11}^{(1)}$ *is singular or ill conditioned, then* **Return**.

Else compute a Gaussian symplectic matrix $Q_2^{-1} \equiv \left[\begin{smallmatrix} I & 0 \\ \Omega & I \end{smallmatrix}\right]$, *with* $\Omega \equiv -\Upsilon_{31}^{(1)} {\Upsilon_{11}^{(1)}}^{-1}$, *so that*

$$Q_2^{-1} \begin{bmatrix} \Upsilon_{11}^{(1)} \\ \Upsilon_{21}^{(1)} \\ \Upsilon_{31}^{(1)} \\ 0 \end{bmatrix} = \begin{bmatrix} \Upsilon_{11}^{(2)} \\ \Upsilon_{21}^{(2)} \\ 0 \\ 0 \end{bmatrix}.$$

Comment: *Since Υ_s is isotropic, it follows that Ω is symmetric. Thus Q_2 is symplectic.*
If $(I + L_{12}\Omega)$ is singular or ill conditioned, then **Return**.
Else set

$$U_2 := \begin{bmatrix} (I + L_{12}\Omega)^{-1} & 0 \\ 0 & I \end{bmatrix}, \quad U_3 := \begin{bmatrix} I & 0 \\ -L_{12}\Omega & I \end{bmatrix},$$

$$Q := QQ_2, \quad U := U_3U_2U,$$

and form

$$N := UNQ = \begin{bmatrix} (I + L_{12}\Omega)^{-1}N_{11} & 0 \\ N_{21} + \Omega - L_{22}\Omega(I + L_{12}\Omega)^{-1}N_{11} & I \end{bmatrix},$$

$$L := ULQ = \begin{bmatrix} I & (I + L_{12}\Omega)^{-1}L_{12} \\ 0 & L_{22} - L_{22}\Omega(I + L_{12}\Omega)^{-1}L_{12} \end{bmatrix}.$$

Comment: *Here the matrix $L_{22} - L_{22}\Omega(I + L_{12}\Omega)^{-1}L_{12} = N_{11}^T(I + \Omega L_{12})^{-1}$ and $(I + L_{12}\Omega)^{-1}L_{12}$ is symmetric.*

Step 3: *Find a symplectic Householder matrix Q_3 such that*

$$Q_3^T \begin{bmatrix} \Upsilon_{11}^{(2)} \\ \Upsilon_{21}^{(2)} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \Upsilon_{11}^{(3)} \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Set

$$N := Q_3^T NQ_3, \quad L := Q_3^T LQ_3, \quad Q := QQ_3, \quad U := Q_3^T U. \qquad \square$$

Remark. (i) This algorithm deflates the maximal (N, L)-stable isotropic subspace of $N - \lambda L$ corresponding to unimodular eigenvalues and gets a reduced symplectic pencil having no unimodular eigenvalue. Consequently, we can use the structure-preserving algorithm proposed by [19] or [29] to compute the stable invariant subspace of the reduced symplectic pencil. (ii) If the matrix $(I + L_{12}\Omega)$ or $\Upsilon_{11}^{(1)}$ in Step 2 is not invertible or ill conditioned, then we return the deflation process to $N - \lambda L$. We deflate the isotropic basis Υ_s from $N - \lambda L$ directly by using symplectic orthogonal transformations and get a reduced symplectic pencil having no unimodular eigenvalue. Then we apply the algorithm of [29] to find the rest of the stable invariant subspace. Although here only orthogonal symplectic transformations are used, it is numerically difficult to keep symplecticity of $N - \lambda L$ explicitly [7]. Hence, it may be numerically troublesome in this case.

5. Numerical examples. In this section we illustrate the numerical performance of our algorithms for a Hamiltonian matrix M. A program based on Algorithms 3.1, 3.3, and 3.6 has been implemented on a SUN 4/470 computer using MATLAB with eps $\approx 10^{-16}$.

Example 5.1. Let

$$A_0 = \operatorname{diag}\left\{[0], J^{(2)}(0)^T, J^{(4)}(0)^T, -I_2\right\},$$
$$H_0 = \operatorname{diag}\left\{[-1], -\Lambda^{(2)}(0), -\Lambda^{(4)}(0), -I_2\right\},$$
$$G_0 = O_{9\times 9},$$

where $J^{(m_j)}(0)$ and $\Lambda^{(m_j)}(0)$ are defined in section 1 with $\Lambda^{(m_j)}(0)(m_j, m_j) = 1$, $j = 1, 2$. It is easily seen that the corresponding Hamiltonian matrix $M_0 = \left[\begin{smallmatrix} A_0 & 0 \\ H_0 & -A_0^T \end{smallmatrix}\right]$ has nonzero eigenvalues -1, -1, 1, 1 and the zero eigenvalue with partial multiplicities $\{2, 4, 8\}$. Now we construct a nontrivial Hamiltonian matrix M by

$$M = \begin{bmatrix} I & V_2 \\ 0 & I \end{bmatrix} \begin{bmatrix} V_1^T & 0 \\ 0 & V_1^{-1} \end{bmatrix} M_0 \begin{bmatrix} V_1^{-T} & 0 \\ 0 & V_1 \end{bmatrix} \begin{bmatrix} I & -V_2 \\ 0 & I \end{bmatrix},$$

where

$$V_1 = \begin{bmatrix} 1 & 1 & & 0 \\ & \ddots & \ddots & \\ & & & 1 \\ 0 & & & 1 \end{bmatrix} \quad \text{and} \quad V_2 = \begin{bmatrix} 1 & 1 & & & & 0 \\ 1 & -1 & 2 & & & \\ & & 2 & 1 & \ddots & \\ & & & \ddots & \ddots & \\ & & & & -1 & 8 \\ 0 & & & & 8 & 1 \end{bmatrix}.$$

The new matrix $M \equiv \left[\begin{smallmatrix} A & G \\ H & -A^T \end{smallmatrix}\right]$ has the same Jordan canonical form as M_0 and has the forms

$$A = \begin{bmatrix} -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 3 & -3 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -3 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & -1 & 1 & 0 & -5 & 5 & 0 & 0 \\ -1 & 1 & -1 & 0 & 1 & 7 & -13 & 6 & 0 \\ -1 & 1 & -1 & 0 & 0 & -4 & 11 & -13 & 7 \\ -1 & 1 & -1 & 0 & 0 & 8 & -14 & 16 & -9 \\ -1 & 1 & -1 & 0 & 0 & 1 & 9 & -15 & 6 \end{bmatrix},$$

$$H = \operatorname{diag}\left\{ \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & -1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} -1 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \right\},$$

$$G = \begin{bmatrix} 2 & -2 & 1 & -3 & 0 & 0 & 0 & 0 & 0 \\ -2 & 10 & -1 & 11 & 4 & 4 & 4 & 4 & 4 \\ 1 & -1 & -3 & -5 & -4 & -1 & -1 & -1 & -1 \\ -3 & 11 & -5 & 9 & 4 & -1 & 3 & 3 & 3 \\ 0 & 4 & -4 & 4 & 17 & -36 & 20 & -40 & -5 \\ 0 & 4 & -1 & -1 & -36 & 75 & -70 & 92 & -53 \\ 0 & 4 & -1 & 3 & 20 & -70 & 98 & -146 & 90 \\ 0 & 4 & -1 & 3 & -40 & 92 & -146 & 178 & -118 \\ 0 & 4 & -1 & 3 & -5 & -53 & 90 & -118 & 115 \end{bmatrix}.$$

It is easy to check that H is negative definite and G is indefinite. The matrix M satisfies the condition of Theorem 5.1 of [8] from H^∞-control problems. Hence, we can apply Algorithms 3.1, 3.3, and 3.6 to find an M-stable Lagrangian subspace $\mathcal{Y}_\mathcal{L}$. We first use Algorithm 3.1 to compute an orthonormal Jordan subbasis $Z[1:14]$ corresponding to zero eigenvalues of M^2. Since zero eigenvalues of M^2 have partial multiplicities $\{1,1,2,2,4,4\}$, we check 2-norms of the following matrices:

	$M^2Z[1:6]$	$M^4Z[7:10]$	$M^6Z[11:12]$	$M^8Z[13:14]$
$\|\cdot\|_2$	3.34e–14	1.72e–12	8.65e–12	1.79e–12

Next, we use Algorithm 3.3 to compute an orthonormal Jordan subbasis $Y[1:14]$ corresponding to a zero eigenvalue of M. The zero eigenvalue of M has partial multiplicities $\{2,4,8\}$; we check 2-norms of the following matrices:

	$MY[1:3]$	$M^2Y[4:6]$	$M^3Y[7:8]$	$M^4Y[9:10]$
$\|\cdot\|_2$	6.44e–14	4.04e–14	8.35e–12	5.27e–13

	$M^5Y[11]$	$M^6Y[12]$	$M^7Y[13]$	$M^8Z[14]$
$\|\cdot\|_2$	2.25e–10	4.22e–12	7.29e–12	2.31e–12

Now, we compute the maximal isotropic subbasis $\Upsilon[1:7] = \Upsilon_s$ of the stable Lagrangian subspace corresponding to zero eigenvalues. At the same time, the isotropicity of $\Upsilon[1:7]$ is checked:

$$\| \Upsilon[1:7]^T J_9 \Upsilon[1:7] \|_2 = 1.96\text{e}-13.$$

Finally, we deflate the zero eigenvalue and the associated subbasis $\Upsilon[1:7]$ of M by using symplectic orthogonal transformations and get a 4×4 Hamiltonian matrix having eigenvalues $\{-1,-1,1,1\}$. Then we use algorithms of [2, 23, 29] to find the rest subbasis $\Upsilon[8:9]$ of the desired M-stable Lagrangian subspace $\mathcal{Y}_\mathcal{L}$. Consequently, a symmetric stable solution X_{sol} of CARE (1.4) is computed by

$$X_{sol} = -\Upsilon[10:18, 1:9]\left(\Upsilon[1:9, 1:9]\right)^{-1}.$$

The 2-norm of the residual of the Riccati equation is $8.71\text{e}-14$. □

6. Conclusions. In this paper, we have presented structure-preserving algorithms for computing an M-stable and an (N,L)-stable Lagrangian subspace of Hamiltonian matrices and symplectic pencils having purely imaginary and unimodular eigenvalues, respectively. These problems often arise in solving the continuous- or discrete-time H^∞-optimal and linear-quadratic control problems, etc. The main approach of our algorithms is to find a maximal isotropic subbasis corresponding to each first half of Jordan blocks of purely imaginary eigenvalues (unimodular eigenvalues, respectively). Furthermore, we deflate the computed isotropic subbasis by using symplectic orthogonal transformations and get a reduced Hamiltonian matrix (symplectic pencil) having no purely imaginary (unimodular) eigenvalues. Then we compute the

maximal stable isotropic subspace of the reduced Hamiltonian matrix (symplectic pencil) by applying some proposed methods of [2, 23, 29]. Thus, we obtain the desired stable Lagrangian subspace by combining these two computed isotropic subspaces. For the continuous case, we first compute an orthonormal Jordan basis corresponding to nonpositive eigenvalues of M^2 and then use it to determine the maximal isotropic Jordan subbasis corresponding to each first half of Jordan blocks of purely imaginary eigenvalues of M. The proposed algorithm is structure preserving and only uses orthogonal transformations. The dominant flops of the algorithm are in the step of reducing M^2 to a skew-Hamiltonian upper triangular matrix. It requires $O(n^2)$ flops for the deflation of the computed isotropic subbasis if the number of purely imaginary eigenvalues is of order 1 compared with the dimension of matrices. Numerical experiments performed on a number of constructive Hamiltonian matrices of dimension 30 with variant sizes of Jordan blocks have shown that our algorithm is stable and reliable in accuracy of the computed maximal isotropic subbasis. For the discrete-time case, we also develop an algorithm to compute the maximal isotropic Jordan subbasis corresponding to each first half of Jordan blocks of unimodular eigenvalues of a symplectic pencil $N - \lambda L$. The approach is analogous to that developed in the continuous case by replacing the M^2-transformation by the $S + S^{-1}$-transformation of the symplectic pencil. The algorithm is structure preserving and uses orthogonal transformations but in the deflation step. Since the algorithm preserves the symplecticity for the pencil type, if the conditions of nonorthogonal transformations in the deflation step are fairly good, the proposed algorithm is still efficient and reliable.

REFERENCES

[1] B. D. O. Anderson and S. Vongpanitlerd, *Network Analysis and Synthesis*, Lecture Notes in Control and Information Science, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[2] G. S. Ammar and V. Mehrmann, *On Hamiltonian and symplectic Hessenberg forms,* Linear Algebra Appl., 149 (1991), pp. 55–72.

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

[4] T. F. Chan, *Rank revealing QR factorizations,* Linear Algebra Appl., 88/89 (1987), pp. 67–82.

[5] D. Clements and K. Glover, *Spectral transformations via Hermitian pencils,* Linear Algebra Appl., 123 (1989), pp. 797–846.

[6] B. A. Francis and J. C. Doyle, *Linear control theory with an H^∞-optimality criterion*, SIAM J. Control Optim., 25 (1987), pp. 815–844.

[7] U. Flaschka, V. Mehrmann, and D. Zywietz, *An analysis of structure preserving methods for symplectic eigenvalue problems,* RAIRO Automat.-Prod. Inform. Ind., 25 (1991), pp. 165–190.

[8] K. Glover, D. J. N. Limebeer, J. C. Doyle, E. M. Kasenally, and M. G. Safonov, *A characterization of all solutions to the fourblock general distance problem*, SIAM J. Control Optim., 39 (1991), pp. 283–324.

[9] G. Golub and C. F. Van Loan, *Matrix Computation,* The Johns Hopkins University Press, Baltimore, MD, 1983.

[10] P. Iglesias, *Robust and Adaptive Control for Discrete-Time Systems*, Ph.D. dissertation, Department of Electrical Engineering, Cambridge University, U.K., 1991.

[11] V. Ionescu and M. Weiss, *Two Riccati formulae for the discrete-time H^∞-control problem*, Internat. J. Control, 57 (1993), pp. 141–195.

[12] C. Kenney, A. J. Laub, and M. Wette, *A stability-enhancing scaling procedure for Schur-Riccati solvers,* Systems Control Lett., 12 (1989), pp. 241–250.

[13] P. Lancaster, A. C. M. Ran, and L. Rodman, *Hermitian solutions of the discrete algebraic Riccati equation,* Internat. J. Control, 44 (1986), pp. 777–802.

[14] P. Lancaster and L. Rodman, *Existence and uniqueness theorems for the algebraic Riccati equation,* Internat. J. Control, 32 (1980), pp. 285–309.

[15] A. J. Laub, *A Schur method for solving algebraic Riccati equations,* IEEE Trans. Automat. Control, 24 (1979), pp. 913–921.

[16] A. J. Laub and K. Meyer, *Canonical forms for symplectic and Hamiltonian Matrices*, Celestial Mech., 9 (1974), pp. 213–238.
[17] W.-W. Lin, *On reducing infinite eigenvalues of regular pencils by a nonequivalence transformation,* Linear Algebra Appl., 78 (1986), pp. 207–231.
[18] W.-W. Lin, *A new method for computing the closed loop eigenvalues of a discrete-time algebraic Riccati equation*, Linear Algebra Appl., 96 (1987), pp. 157–180.
[19] L.-Z. Lu and W.-W Lin, *An iterative algorithm for the solution of the discrete-time algebraic Riccati equation*, Linear Algebra Appl., 188/189 (1993), pp. 465–488.
[20] V. Mehrmann, *The Autonomous Linear Quadratic Control Problem,* Springer-Verlag, Berlin, 1991.
[21] C. Paige and C. F. Van Loan, *A Schur decomposition for Hamiltonian matrices*, Linear Algebra Appl., 41 (1981), pp. 11–32.
[22] R. V. Patel, *On computing the eigenvalues of a symplectic pencils*, Linear Algebra Appl., 188 (1993), pp. 591–611.
[23] R. V. Patel, Z. Lin, and P. Misra, *Computation of stable invariant subspaces of Hamiltonian matrices*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 284–298.
[24] P. H. Petkov, N. D. Christov, and M. M. Konstantinov, *On the numerical properties of the Schur approach for solving the matrix Riccati equation*, Systems Control Lett., 9 (1987), pp. 197–201.
[25] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
[26] A. J. Van Der Schaft and J. C. Willems, *A new procedure for stochastic realization of spectral density matrices*, SIAM J. Control Optim., 22 (1984), pp. 845–855.
[27] C. F. Van Loan, *A symplectic method for approximating all the eigenvalues of a Hamiltonian matrix*, Linear Algebra Appl., 61 (1984), pp. 233–251.
[28] J. C. Willems, *Least squares stationary optimal control and the algebraic Riccati equations*, IEEE Trans. Automat. Control, 16 (1971), pp. 621–634.
[29] H.-G. Xu, *Solving Algebraic Riccati Equations via Skew-Hamiltonian Matrices,* Ph.D. thesis, Department of Math, Fudan University, Shanghai, China, 1991.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 3, pp. 615–632, July 1997

006

THE MATRIX SIGN FUNCTION METHOD AND THE COMPUTATION OF INVARIANT SUBSPACES*

RALPH BYERS†, CHUNYANG HE‡, AND VOLKER MEHRMANN§

Abstract. A perturbation analysis shows that if a numerically stable procedure is used to compute the matrix sign function, then it is competitive with conventional methods for computing invariant subspaces. Stability analysis of the Newton iteration improves an earlier result of Byers and confirms that ill-conditioned iterates may cause numerical instability. Numerical examples demonstrate the theoretical results.

Key words. matrix sign function, invariant subspaces, perturbation theory

AMS subject classifications. 65F15, 65G99

PII. S0895479894277454

1. Introduction.

If $A \in \mathbf{R}^{n\times n}$ has no eigenvalue on the imaginary axis, then the matrix sign function $\operatorname{sign}(A)$ may be defined as

$$\operatorname{sign}(A) = \frac{1}{\pi i}\int_\gamma (zI - A)^{-1}dz - I, \tag{1}$$

where γ is any simple closed curve in the complex plane enclosing all eigenvalues of A with positive real part. The sign function is used to compute eigenvalues and invariant subspaces [2, 4, 6, 13, 14] and to solve Riccati and Sylvester equations [9, 15, 16, 28]. The matrix sign function is attractive for machine computation because it can be efficiently evaluated by relatively simple numerical methods. Some of these are surveyed in [28]. It is particularly attractive for large dense problems to be solved on computers with advanced architectures [2, 11, 16, 33].

Beavers and Denman use the following equivalent definition [6, 13]. Let $A = XJX^{-1}$ be the Jordan canonical decomposition of a matrix A having no eigenvalues on the imaginary axis. Let the diagonal part of J be given by the matrix $D = \operatorname{diag}(d_1, \ldots, d_n)$. If $S = \operatorname{diag}(s_1, \ldots, s_n)$, where

$$s_i = \begin{cases} +1 & \text{if } \Re(d_i) > 0, \\ -1 & \text{if } \Re(d_i) < 0, \end{cases}$$

then $\operatorname{sign}(A) = XSX^{-1}$.

Let $\mathcal{V}^+ = \mathcal{V}^+(A)$ be the invariant subspace of A corresponding to eigenvalues with positive real part, let $\mathcal{V}^- = \mathcal{V}^-(A)$ be the invariant subspace of A corresponding to eigenvalues with negative real part, let $P^+(A) = P^+$ be the skew projection onto $\mathcal{V}^+$ parallel to $\mathcal{V}^-$, and let $P^- = P^-(A)$ be the skew projection onto $\mathcal{V}^-$ parallel to

* Received by the editors November 21, 1994; accepted for publication (in revised form) by P. Van Dooren July 12, 1996.

http://www.siam.org/journals/simax/18-3/27745.html

† University of Kansas, Department of Mathematics, Lawrence, KS 66045 (byers@ariel.math.ukans.edu). Partial support was received from National Science Foundation grants INT-8922444 and CCR-9404425 and University of Kansas GRF allocation 3514-20-0038.

‡ University of Kansas, Department of Mathematics, Lawrence, KS 66045 (he@math.ukans.edu).

§ TU-Chemnitz-Zwickau, Fak. f. Mathematik, D-09107 Chemnitz, Germany (mehrmann@mathematik.tu-chemnitz.de). Partial support was received from Deutsche Forschungsgemeinschaft, Projekt La 767/3-2.

$\mathcal{V}^+$. Using the same contour γ as in (1), the projection P^+ has the resolvent integral representation [23, p. 67], [2]

$$P^+ = \frac{1}{2\pi i}\int_\gamma (zI - A)^{-1} dz. \tag{2}$$

It follows from (1) and (2) that $\operatorname{sign}(A) = P^+ - P^- = 2P^+ - I = I - 2P^-$.

The matrix sign function was introduced using definition (1) by Roberts in a 1971 technical report [34] which was not published until 1980 [35]. Kato [23, p. 67] reports that the resolvent integral (2) goes back to 1946 [12] and 1949 [21, 22].

There is some concern about the numerical stability of numerical methods based upon the matrix sign function [2, 8, 19]. In this paper, we demonstrate that evaluating the matrix sign function is a more ill-conditioned computational problem than the problem of finding bases of the invariant subspaces $\mathcal{V}^+$ and $\mathcal{V}^-$. See Example 1 in section 3. Nevertheless, we also give perturbation and error analyses, which show that (at least for Newton's method for the computation of the matrix sign function [8, 9]) *in most circumstances* the accuracy is competitive with conventional methods for computing invariant subspaces. Our analysis improves some of the perturbation bounds in [2, 8, 18, 24].

In section 2 we establish some notation and clarify the relationship between the matrix sign function and the Schur decomposition. The next two sections give a perturbation analysis of the matrix sign function and its invariant subspaces. Section 5 gives a posteriori bounds on the forward and backward error associated with a corrupted value of $\operatorname{sign}(S)$. Section 6 contains a stability analysis of the Newton iteration.

Throughout the paper, $\|\cdot\|$ denotes the spectral norm, $\|\cdot\|_1$ the 1-norm (or column sum norm), and $\|\cdot\|_F$ the Frobenius norm $\|\cdot\|_F = \sqrt{\sum |a_{ij}|^2}$. The set of eigenvalues of a matrix A is denoted by $\lambda(A)$. The open left half-plane is denoted by $\mathbf{C}^-$ and the open right half-plane is denoted by $\mathbf{C}^+$. Borrowing some terminology from engineering, we refer to the invariant subspace $\mathcal{V}^- = \mathcal{V}^-(A)$ of a matrix $A \in \mathbf{R}^{n\times n}$ corresponding to eigenvalues in $\mathbf{C}^-$ as the *stable invariant subspace* and the subspace $\mathcal{V}^+ = \mathcal{V}^+(A)$ corresponding to eigenvalues in $\mathbf{C}^+$ as the *unstable invariant subspace*. We use $P^+ = P^+(A)$ for the skew projection onto $\mathcal{V}^+$ parallel to $\mathcal{V}^-$ and $P^- = P^-(A)$ for the skew projection onto $\mathcal{V}^-$ parallel to $\mathcal{V}^+$.

2. Relationship with the Schur decomposition. Suppose that A has the Schur form

$$Q^H A Q = \begin{array}{c} \\ k \\ n-k \end{array} \begin{array}{c} \begin{array}{cc} k & n-k \end{array} \\ \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \end{array}, \tag{3}$$

where $\lambda(A_{11}) \subset \mathbf{C}^-$ and $\lambda(A_{22}) \subset \mathbf{C}^+$ [17]. If Y is a solution of the Sylvester equation

$$YA_{22} - A_{11}Y = 2A_{12}, \tag{4}$$

then

$$Q^H \operatorname{sign}(A) Q = \begin{array}{c} \\ k \\ n-k \end{array} \begin{array}{c} \begin{array}{cc} k & n-k \end{array} \\ \begin{bmatrix} -I & Y \\ 0 & I \end{bmatrix} \end{array}, \tag{5}$$

$$Q^H P^- Q = \begin{matrix} \\ k \\ n-k \end{matrix} \begin{matrix} \begin{matrix} k & n-k \end{matrix} \\ \begin{bmatrix} I & -\frac{1}{2}Y \\ 0 & 0 \end{bmatrix} \end{matrix},$$

and

$$Q^H P^+ Q = \begin{matrix} \\ k \\ n-k \end{matrix} \begin{matrix} \begin{matrix} k & n-k \end{matrix} \\ \begin{bmatrix} 0 & \frac{1}{2}Y \\ 0 & I \end{bmatrix} \end{matrix}.$$

The solution of (4) has the integral representation

$$Y = \frac{1}{\pi i} \int_\gamma (zI - A_{11})^{-1} A_{12} (zI - A_{22})^{-1} dz, \tag{6}$$

where γ is a closed contour containing all eigenvalues of A with positive real part [29, 36].

The stable invariant subspace of A is the range (or column space) of $\operatorname{sign}(A) - I = -2P^-$. If

$$(\operatorname{sign}(A) - I)\Pi = QR = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 & R_2 \\ 0 & 0 \end{bmatrix} \tag{7}$$

is a QR factorization with column pivoting [1, 17], where Q and R are partitioned in the obvious way, then the columns of Q_1 form an orthonormal basis of this subspace. Here Q is orthogonal, Π is a permutation matrix, R is upper triangular, and R_1 is nonsingular.

It is not difficult to use the singular value decomposition of Y to show that

$$\|\operatorname{sign}(A)\| = \frac{1}{2}\|Y\| + \sqrt{1 + \frac{1}{4}\|Y\|^2}. \tag{8}$$

It follows from (4) that

$$\|Y\| \le \frac{2\|A_{12}\|}{\operatorname{sep}(A_{11}, A_{22})}, \tag{9}$$

where sep is defined as in [17] by $\operatorname{sep}(A_{11}, A_{22}) = \min_{Z \neq 0} \frac{\|A_{11}Z - ZA_{22}\|_F}{\|Z\|_F}$.

3. The effect of backward errors. In this section we discuss the sensitivity of the matrix sign function subject to perturbations. For a perturbation matrix E, we give first-order estimates for $\operatorname{sign}(A + E)$ in terms of submatrices and powers of $\|E\|$.

Based on Fréchet derivatives, Kenney and Laub [24] presented a first-order perturbation theory for the matrix sign function via the solution of a Sylvester equation. Mathias [30] derived an expression for the Fréchet derivative using the Schur decomposition. Kato's encyclopedic monograph [23] includes an extensive study of series representations and perturbation bounds for eigenprojections. In this section we derive an expression for the Fréchet derivative using integral formulas.

Let

$$d_A = \min_{\mu \in \mathbf{R}} \sigma_{min}(A - \mu i I),$$

where $\sigma_{min}(A - \mu i I)$ is the smallest singular value of $A - \mu i I$. The quantity d_A is the distance from A to the nearest complex matrix with an eigenvalue on the imaginary

axis. Practical numerical techniques for calculating d_A appear in [7, 10]. If $\|E\| < d_A$, then E is too small to perturb an eigenvalue of A on or across the imaginary axis. It follows that for $\|E\| < d_A$, $\operatorname{sign}(A+E)$ and the stable and unstable invariant subspaces of $A+E$ are smooth functions of E.

Consider the relatively simple case in which A is block diagonal.

LEMMA 3.1. *Suppose A is block diagonal,*

$$A = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix},$$

where $\lambda(A_{11}) \subset \mathbf{C}^-$ and $\lambda(A_{22}) \subset \mathbf{C}^+$. Partition the perturbation $E \in \mathbf{R}^{n\times n}$ conformally with A as

$$E = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}. \tag{10}$$

If $\|E\| < d_A$, then

$$\operatorname{sign}(A+E) = \operatorname{sign}(A) + 2\left(\begin{bmatrix} 0 & F_{12} \\ F_{21} & 0 \end{bmatrix}\right) + O(\|E\|^2),$$

where F_{12} and F_{21} satisfy the Sylvester equations

$$A_{22}F_{21} - F_{21}A_{11} = E_{21}, \tag{11}$$

$$F_{12}A_{22} - A_{11}F_{12} = E_{12}. \tag{12}$$

Proof. The hypothesis that $\|E\| < d_A$ implies that the eigenvalues of $A_{11}+E_{11}$ have negative real part and the eigenvalues of $A_{22}+E_{22}$ have positive real part. In the definition (1) choose the contour γ to enclose $\lambda(A_{22})$ and $\lambda(A_{22}+E_{22})$ but neither $\lambda(A_{11})$ nor $\lambda(A_{11}+E_{11})$. In particular, for all complex numbers z lying on the contour γ, $zI-A$ and $zI-(A+E)$ are nonsingular and

$$\begin{aligned}(zI-(A+E))^{-1} = (zI-A)^{-1} &+ (zI-A)^{-1}E(zI-A)^{-1} \\ &+ (zI-A)^{-1}E(zI-A)^{-1}E(zI-(A+E))^{-1}.\end{aligned}$$

So,

$$\begin{aligned}\operatorname{sign}(A+E) &= \frac{1}{\pi i}\int_\gamma (zI-(A+E))^{-1}\,dz - I \\ &= \frac{1}{\pi i}\int_\gamma ((zI-A)^{-1} + (zI-A)^{-1}E(zI-A)^{-1})\,dz - I \\ &\quad + O(\|E\|^2) \\ &= \operatorname{sign}(A) + 2F + O(\|E\|^2),\end{aligned}$$

where

$$F = \frac{1}{2\pi i}\int_\gamma (zI-A)^{-1}E(zI-A)^{-1}\,dz.$$

Partitioning F conformally with E and A, then we have

$$F_{11} = \frac{1}{2\pi i}\int_\gamma (zI-A_{11})^{-1}E_{11}(zI-A_{11})^{-1}\,dz,$$

$$F_{12} = \frac{1}{2\pi i}\int_\gamma (zI - A_{11})^{-1}E_{12}(zI - A_{22})^{-1}\,dz,$$
$$F_{21} = \frac{1}{2\pi i}\int_\gamma (zI - A_{22})^{-1}E_{21}(zI - A_{11})^{-1}\,dz,$$
$$F_{22} = \frac{1}{2\pi i}\int_\gamma (zI - A_{22})^{-1}E_{22}(zI - A_{22})^{-1}\,dz.$$

As in (6), F_{12} and F_{21} are the solutions to the Sylvester equations (11) and (12) [29, 36]. The contour γ encloses no eigenvalues of A_{11}, so $(zI-A_{11})^{-1}E_{11}(zI-A_{11})^{-1}$ is analytic inside γ and $F_{11} = 0$.

We first prove that $F_{22} = 0$ in the case in which A_{22} is diagonalizable, say $A_{22} = X\Lambda X^{-1}$, where $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{n-k})$. Then

$$F_{22} = X\left(\frac{1}{2\pi i}\int_\gamma (zI - \Lambda)^{-1}(X^{-1}E_{22}X)(zI - \Lambda)^{-1}\,dz\right)X^{-1}.$$

Each component of the above integral is of the form $\int_\gamma c(z - \lambda_j)^{-1}(z - \lambda_k)^{-1}\,dz$ for some constant c. If $j = k$ then this is the integral of a residue-free holomorphic function and hence it vanishes. If $j \neq k$, then

$$\int_\gamma \frac{c}{(z-\lambda_i)(z-\lambda_j)}\,dz = \int_\gamma \frac{c}{\lambda_i - \lambda_j}\left(\frac{1}{z-\lambda_i} - \frac{1}{z-\lambda_j}\right)dz = 0.$$

The general case follows by taking limits of the diagonalizable case and using the dominated convergence theorem. □

The following theorem gives the general case.

THEOREM 3.2. *Let the Schur form of A be given by* (3) *and partition E conformally as*

$$Q^H E Q = \begin{array}{c} \\ k \\ n-k \end{array}\begin{array}{c} \begin{array}{cc} k & n-k \end{array} \\ \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} \end{array}. \tag{13}$$

If $\|SES^{-1}\| < d_A$, where

$$S = \begin{bmatrix} I & -\frac{Y}{2} \\ 0 & I \end{bmatrix},$$

and Y satisfies (4)*, then*

$$\operatorname{sign}(A + E) = \operatorname{sign}(A) + E_t - \operatorname{sign}(A)E_p\operatorname{sign}(A) + O(\|E\|^2),$$

where

$$E_t = Q\begin{bmatrix} 0 & 2\tilde{E}_{12} + \frac{Y\tilde{E}_{21}Y}{2} \\ \tilde{E}_{21} & 0 \end{bmatrix}Q^H,$$
$$E_p = Q\begin{bmatrix} 0 & 0 \\ \tilde{E}_{21} & 0 \end{bmatrix}Q^H,$$

$\tilde{E}_{21}$ satisfies the Sylvester equation

$$A_{22}\tilde{E}_{21} - \tilde{E}_{21}A_{11} = E_{21}, \tag{14}$$

and $\tilde{E}_{12}$ *satisfies*

$$\tilde{E}_{12}A_{22} - A_{11}\tilde{E}_{12} = E_{12} - \frac{YE_{22}}{2} + \frac{E_{11}Y}{2} - \frac{YE_{21}Y}{4}. \tag{15}$$

Proof. If $S = \left[\begin{smallmatrix} I & -\frac{Y}{2} \\ 0 & I \end{smallmatrix}\right]$, then

$$S \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} S^{-1} = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}$$

and

$$S \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} S^{-1} = \begin{bmatrix} E_{11} - \frac{YE_{21}}{2} & E_{12} - \frac{YE_{22}}{2} + \frac{E_{11}Y}{2} - \frac{YE_{21}Y}{4} \\ E_{21} & \frac{E_{21}Y}{2} + E_{22} \end{bmatrix}.$$

It follows from Lemma 3.1 that

$$\operatorname{sign}(SQ^H(A+E)QS^{-1}) = \begin{bmatrix} -I & 0 \\ 0 & I \end{bmatrix} + 2\begin{bmatrix} 0 & \tilde{E}_{12} \\ \tilde{E}_{21} & 0 \end{bmatrix} + O(\|E\|^2).$$

Since $\operatorname{sign}(SAS^{-1}) = S\operatorname{sign}(A)S^{-1}$, multiplying QS^{-1} on the left side and SQ^H on the right side of the above equation, we have

$$\operatorname{sign}(A+E) = \operatorname{sign}(A) + Q \begin{bmatrix} Y\tilde{E}_{21} & 2\tilde{E}_{12} - \frac{Y\tilde{E}_{21}Y}{2} \\ 2\tilde{E}_{21} & -\tilde{E}_{21}Y \end{bmatrix} Q^H + O(\|E\|^2). \tag{16}$$

It is easy to verify that

$$\begin{aligned} &\begin{bmatrix} Y\tilde{E}_{21} & 2\tilde{E}_{12} - \frac{Y\tilde{E}_{21}Y}{2} \\ 2\tilde{E}_{21} & -\tilde{E}_{21}Y \end{bmatrix} \\ &\quad = \begin{bmatrix} 0 & 2\tilde{E}_{12} + \frac{Y\tilde{E}_{21}Y}{2} \\ \tilde{E}_{21} & 0 \end{bmatrix} - \begin{bmatrix} -I & Y \\ 0 & I \end{bmatrix} \begin{bmatrix} 0 & 0 \\ \tilde{E}_{21} & 0 \end{bmatrix} \begin{bmatrix} -I & Y \\ 0 & I \end{bmatrix}. \end{aligned} \tag{17}$$

The theorem follows from

$$Q^H \operatorname{sign}(A) Q = \operatorname{sign}(Q^H A Q) = \begin{bmatrix} -I & Y \\ 0 & I \end{bmatrix}. \qquad \square$$

If d_A is small relative to $\|A\|$ or $\|Y\|$ is large relative to $\|A\|$, then the hypothesis that $\|SES^{-1}\| < d_A$ may be restrictive. However, a small value of d_A indicates that A is very near a discontinuity in $\operatorname{sign}(A)$. A large value of $\|Y\|$ indicates that $\operatorname{sep}(A_{11}, A_{22})$ is small and the stable invariant subspace is ill conditioned [37, 39].

Of course Theorem 3.2 also gives first-order perturbations for the projections $P^+ = P^+(A)$ and $P^- = P^-(A)$.

COROLLARY 3.3. *Let the Schur form of* A *be given as in* (3) *and let* E *be as in* (13). *Under the hypothesis of Theorem* 3.2, *the projections* P^+ *and* P^- *satisfy*

$$\begin{aligned} &P^{\pm}(A+E) \\ &= P^{\pm}(A) + \frac{1}{2}\left(E_t - \operatorname{sign}(A)E_p\operatorname{sign}(A)\right) + O(\|E\|^2) \\ &= P^{\pm}(A) + \frac{1}{2}\left(E_t - \left(P^{\pm}(A) - P^{\mp}(A)\right)E_p\left(P^{\pm}(A) - P^{\mp}(A)\right)\right) \\ &\quad + O(\|E\|^2) \\ &= P^{\pm}(A) + \frac{1}{2}\left(E_t - \left(2P^{\pm}(A) - I\right)E_p\left(2P^{\pm}(A) - I\right)\right) + O(\|E\|^2), \end{aligned}$$

where E_t and E_p are as in the statement of Theorem 3.2.

Taking norms in Theorem 3.2 gives the first-order perturbation bounds of the next corollary.

COROLLARY 3.4. *Let the Schur form of A be given as in* (3) *and E as in* (13) *and let $0 < \delta = \min\{\mathrm{sep}(A_{11}, A_{22}), \mathrm{sep}(A_{22}, A_{11})\}$. Then the first-order perturbation of the matrix sign function stated in Theorem* 3.2 *is bounded by*

$$\|\operatorname{sign}(A+E) - \operatorname{sign}(A)\| \leq \frac{4}{\delta}\left(1 + \frac{\|A_{12}\|}{\delta}\right)^2 \|E\| + O(\|E\|^2).$$

On first examination, Corollary 3.4 is discouraging. It suggests that calculating the matrix sign function may be more ill conditioned than finding bases of the stable and unstable invariant subspace. If the matrix A whose Schur decomposition appears in (3) is perturbed to $A + E$, then the stable invariant subspace, range(Q_1), is perturbed to range$(Q_1 + Q_2 W)$, where $\|W\| \leq 2\|E\|/\delta$ [37, 39]. Corollary 3.4 and the following example show that $\operatorname{sign}(A + E)$ may indeed differ from $\operatorname{sign}(A)$ by a factor of δ^{-3} which may be much larger than $\|E\|/\delta$.

Example 1. Let

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

The matrix A is already in Schur form, so $\mathrm{sep}(A_{11}, A_{22}) = 2\eta$. If $\epsilon < \eta < 1$, then we have

$$\operatorname{sign}(A) = \begin{bmatrix} -1 & \eta^{-1} \\ 0 & 1 \end{bmatrix},$$
$$\operatorname{sign}(A+E) = \frac{1}{\sqrt{\eta^2+\epsilon}} \begin{bmatrix} -\eta & 1 \\ \epsilon & \eta \end{bmatrix}.$$

The difference is

$$\operatorname{sign}(A+E) - \operatorname{sign}(A) = \epsilon \begin{bmatrix} \eta^{-2}/2 & -\eta^{-3}/2 \\ \eta^{-1} & -\eta^{-2}/2 \end{bmatrix} + O(\epsilon^2).$$

Perturbing A to $A + E$ does indeed perturb the matrix sign function by a factor of δ^{-3}.

Of course there is no rounding error in Example 1, so the stable invariant subspace of $A + E$ is also the stable invariant subspace of $\operatorname{sign}(A + E)$ and, in particular, evaluating $\operatorname{sign}(A + E)$ exactly has done no more damage than perturbing A. The stable invariant subspace of A is $\mathcal{V}^-(A) = \mathrm{range}([\begin{smallmatrix}1\\0\end{smallmatrix}])$; the stable invariant subspace of $A + E$ and $\operatorname{sign}(A + E)$ is

$$\mathcal{V}^-(A+E) = \mathrm{range}\left(\begin{bmatrix} 1 \\ \frac{-\epsilon}{\eta+\sqrt{\eta^2+\epsilon}} \end{bmatrix}\right) = \mathrm{range}\left(\begin{bmatrix} 1 \\ \frac{-\epsilon}{2\eta} \end{bmatrix}\right) + O(\epsilon^2).$$

For a general small perturbation matrix E, the angle between $\mathcal{V}^-(A)$ and $\mathcal{V}^-(A+E)$ is of order no larger than $O(1/\eta)$ [17, 37, 39]. *The matrix sign function (and the projections P^- and P^+) may be significantly more ill conditioned than the stable and*

unstable invariant subspaces. Nevertheless, we argue in this paper that despite the possible poor conditioning of the matrix sign function, the invariant subspaces are usually preserved about as accurately as their native conditioning permits.

However, if the perturbation E is large enough to perturb an eigenvalue across or on the imaginary axis, then the stable and unstable invariant subspaces may become confused and cannot be extracted from $\operatorname{sign}(A+E)$. This may occur even when the invariant subspaces are well conditioned, since the sign function is not defined in this case. In geometric terms, in this situation A is within distance $\|E\|$ of a matrix with an eigenvalue with zero real part. This is a fundamental difficulty of any method that identifies the two invariant subspaces by the signs of the real parts of the corresponding eigenvalues of A.

4. Perturbation theory for invariant subspaces of the matrix sign function. In this section we discuss the accuracy of the computation of the stable invariant subspace of A via the matrix sign function.

An easy first observation is that if the computed value of $\operatorname{sign}(A)$ is the exact value of $\operatorname{sign}(A+E)$ for some perturbation matrix E, then the exact stable invariant subspace of $\operatorname{sign}(A+E)$ is also an invariant subspace of $A+E$. Let A have Schur form (3) and let E be a perturbation matrix partitioned conformally as in (13). Let Q_1 be the first k columns of Q and Q_2 be the remaining $n-k$ columns. If

$$0 \le \frac{\|E_{21}\|\,(\|A_{12}\| + \|E_{12}\|)}{\operatorname{sep}(A_{11}, A_{22}) - \|E_{11}\| - \|E_{22}\|} < \frac{1}{4},$$

then A has a stable invariant subspace $\mathcal{V}^-(A) = \operatorname{range}(Q_1)$ and $A+E$ has an invariant subspace $\operatorname{range}(Q_1 + Q_2 W)$, where W satisfies

$$\|W\| \le \frac{2\|E_{21}\|}{\operatorname{sep}(A_{11}, A_{22}) - \|E_{11}\| - \|E_{22}\|} \tag{18}$$

$$\le 2\epsilon_{21} \frac{\|A\|}{\operatorname{sep}(A_{11}, A_{22})} + O(\epsilon^2), \tag{19}$$

where $\epsilon_{21} = \|E_{21}\|/\|A\|$ [17, 37, 39]. The singular values of W are the tangents of the canonical angles between $\mathcal{V}^-(A) = \operatorname{range}(Q_1)$ and $\operatorname{range}(Q_1 + Q_2 W)$. In particular, the canonical angles are at most of order $O(1/\operatorname{sep}(A_{11}, A_{22}))$.

Unfortunately, in general, we cannot apply backward error analysis; i.e., we cannot guarantee that the computed value of $\operatorname{sign}(A)$ is exactly the value of $\operatorname{sign}(A+E)$ for some perturbation E. Consider instead the effect of forward errors. Let $B = \operatorname{sign}(A) + F$, where F represents the forward error in evaluating the matrix sign function. Let A have Schur form (3). Partition $Q^H \operatorname{sign}(A) Q$ and $Q^H F Q$ as

$$Q^H \operatorname{sign}(A) Q = \begin{matrix} \\ k \\ n-k \end{matrix} \begin{matrix} \begin{matrix} k & n-k \end{matrix} \\ \begin{bmatrix} -I & Y \\ 0 & I \end{bmatrix} \end{matrix}$$

and

$$Q^H F Q = \begin{matrix} \\ k \\ n-k \end{matrix} \begin{matrix} \begin{matrix} k & n-k \end{matrix} \\ \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix} \end{matrix},$$

where Q is the unitary factor from the Schur decomposition of A (3) and Y is a solution of (4).

Assume that

$$0 \le \frac{\|F_{21}\|(\|Y\| + \|F_{12}\|)}{\operatorname{sep}(I,-I) - \|F_{11}\| - \|F_{22}\|} < \frac{1}{4},$$

and let $\phi_{21} = \|F_{21}\|/\|\operatorname{sign}(A)\|$.

Perturbing $\operatorname{sign}(A)$ to $\operatorname{sign}(A) + F$ changes the invariant subspace $\mathcal{V}^-(A) = \mathcal{V}^-(\operatorname{sign}(A)) = \operatorname{range}(Q_1)$ to $\operatorname{range}(Q_1 + Q_2 W_s)$, where [17, 37, 39]

$$\|W_s\| \le 2\|F_{21}\|/(2 - \|F_{11}\| - \|F_{22}\|),$$

and by (8) and (9)

$$\begin{aligned}\|F_{21}\| &= \phi_{21}\|\operatorname{sign}(A)\| \\ &\le \phi_{21}\left(\frac{1}{2}\|Y\| + \sqrt{1 + \frac{\|Y\|^2}{4}}\right) \\ &\le \phi_{21}\left(2\frac{\|A_{12}\|}{\operatorname{sep}(A_{11}, A_{22})} + 1\right).\end{aligned}$$

Since $\operatorname{sep}(A_{11}, A_{22}) \le 2\|A\|_F$, W_s obeys the bounds

$$\begin{aligned}\|W_s\| &\le 2\phi_{21}\frac{2\frac{\|A_{12}\|}{\operatorname{sep}(A_{11},A_{22})} + 1}{2 - \|F_{11}\| - \|F_{22}\|} \\ &\le 4\phi_{21}\left(\frac{\|A\|_F}{\operatorname{sep}(A_{11}, A_{22})}\right) + O\left(\frac{\|F\|^2}{\|\operatorname{sign}(A)\|^2}\right). \qquad (20)\end{aligned}$$

Comparing (19) with (20), we see that the error bound (20) is no greater than twice the error bound (19). Loosely speaking, a small relative error in $\operatorname{sign}(A)$ of size ϵ might perturb the stable invariant subspace by not much more than twice as much as a relative error of size ϵ in A can.

Therefore, the stable and unstable invariant subspaces of $\operatorname{sign}(A)$ may be less ill conditioned and are never significantly more ill conditioned than the corresponding invariant subspaces of A. There is no fundamental numerical instability in evaluating the matrix sign function as a means of extracting invariant subspaces. However, numerical methods used to evaluate the matrix sign function may or may not be numerically unstable.

Example 1 *continued.* To illustrate the results, we give a comparison of our perturbation bounds and the bounds given in [3] for both the matrix sign function and the invariant subspaces in the case of Example 1. The distance to the nearest ill-posed problem, i.e., $d_A = \min_\mu \sigma_{\mu \in \mathbf{R}}(A - \mu i I)$, where $\sigma_{min}(A - \mu i I)$ is the smallest singular value of $(A - \mu i I)$, leads to an overestimation of the error in [3]. Since $d_A \approx \eta^{-2}$, the bounds given in [3] are, respectively, $O(\eta^{-4})$ for the matrix sign function and $O(\eta^{-2})$ for the invariant subspaces.

5. A posteriori backward and forward error bounds. A priori backward and forward error bounds for the evaluation of the matrix sign function remain elusive. However, it is not difficult to derive a posteriori error bounds for both backward and forward error.

We will need the following lemma to estimate the distance between a matrix S and $\operatorname{sign}(S)$.

LEMMA 5.1. *If $S \in \mathbf{R}^{n\times n}$ has no eigenvalue with zero real part and $\|\operatorname{sign}(S)S^{-1} - I\| < 1$, then $\|\operatorname{sign}(S) - S\| \le \|S^{-1} - S\|$.*

Proof. Let $F = \operatorname{sign}(S) - S$. The matrices F, S, and $\operatorname{sign}(S)$ commute, so

$$I = \operatorname{sign}(S)^2 = (S + F)^2 = S^2 + 2SF + F^2.$$

This implies that

$$\frac{S^{-1} - S}{2} - \frac{S^{-1}F^2}{2} = F.$$

Taking norms and using $\|FS^{-1}\| = \|\operatorname{sign}(S)S^{-1} - I\| < 1$, we get

$$\frac{1}{2}\|S^{-1} - S\| + \frac{1}{2}\|F\| \ge \|F\|$$

and the lemma follows. ∎

It is clear from the proof of Lemma 5.1 that $(\operatorname{sign}(S) - S) \approx (S^{-1} - S)/2$ is asymptotically correct as $\|\operatorname{sign}(S) - S\|$ tends to zero. The bound in the lemma tends to overestimate smaller values of $\|\operatorname{sign}(S) - S\|$ by a factor of two.

Suppose that a numerical procedure for evaluating $\operatorname{sign}(A)$ applied to a matrix $A \in \mathbf{R}^{n\times n}$ produces an approximation $S \in \mathbf{R}^{n\times n}$. Consider the problem of finding small norm solutions $E \in \mathbf{R}^{n\times n}$ and $F \in \mathbf{R}^{n\times n}$ to $\operatorname{sign}(A+E) = S+F$. Of course, this does not uniquely determine E and F. Common algorithms for evaluating $\operatorname{sign}(A)$, such as Newton's method for the square root of I, guarantee that S is very nearly a square root of I [19]; i.e., S is a close approximation of $\operatorname{sign}(S)$. In the following theorem, we have arbitrarily taken $F = \operatorname{sign}(S) - S$.

THEOREM 5.2. *If $\|\operatorname{sign}(S)S^{-1} - I\| < 1$ and $\|\operatorname{sign}(S)A - A\operatorname{sign}(S)\| < d_A$, then $\operatorname{sign}(A + E) = S + F$ for perturbation matrices E and F satisfying*

$$\|F\| \le \|S^{-1} - S\|$$

and

$$\frac{\|E\|}{\|A\|} \le \frac{\|\operatorname{sign}(S)A - A\operatorname{sign}(S)\|}{2\|A\|} \tag{21}$$

$$\le \frac{\|SA - AS\|}{2\|A\|} + \|S^{-1} - S\|. \tag{22}$$

Proof. The matrices $S+F$ and $A+E$ commute, so an underdetermined, consistent system of equations for E in terms of S, A, and $F = \operatorname{sign}(S) - S$ is

$$E(S + F) - (S + F)E = \operatorname{sign}(S)A - A\operatorname{sign}(S) = (SA - AS) + (FA - AF). \tag{23}$$

Let

$$U^H \operatorname{sign}(S)U = \begin{bmatrix} -I & Y \\ 0 & I \end{bmatrix} \tag{24}$$

be a Schur decomposition of $\operatorname{sign}(S)$ whose unitary factor is U and whose triangular factor is on the right-hand side of (24). Partition U^HEU and U^HAU conformally with the right-hand side of (24) as

$$U^HEU = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}$$

and

$$U^H A U = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

Multiplying (23) on the left by U^H and on the right by U and partitioning gives

$$\begin{bmatrix} -YE_{21} & E_{11}Y - YE_{22} + 2E_{12} \\ -2E_{21} & E_{21}Y \end{bmatrix} = \begin{bmatrix} YA_{21} & -A_{11}Y + YA_{22} - 2A_{12} \\ 2A_{21} & -A_{21}Y \end{bmatrix}.$$

One of the infinitely many solutions for E is given by

$$U^H E U = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2}(-A_{11}Y + YA_{22} - 2A_{12}) \\ -A_{21} & 0 \end{bmatrix}. \tag{25}$$

For this choice of E, we have

$$\begin{aligned} 2\|E\| &\le \|\operatorname{sign}(S)A - A\operatorname{sign}(S)\| \\ &\le \|SA - AS\| + \|FA - AF\| \\ &\le \|SA - AS\| + 2\|S^{-1} - S\|\,\|A\|, \end{aligned}$$

from which the theorem follows. □

Lemma 5.1 and Theorem 5.2 agree well with intuition. To assure small forward error, S must be a good approximate square root of I and, in addition, to assure small backward error, $\operatorname{sign}(S)$ must nearly commute with the original data matrix A. Newton's method for a root of $X^2 - I$ tends to do a good job of both [19]. (Note that, in general, Newton's method makes a poor algorithm to find a square root of a matrix. The square root of I is a special case. See [19] for details.) In particular, the hypothesis that $\|\operatorname{sign}(S)S^{-1} - I\| < 1$ is usually satisfied when the matrix sign function is computed by the Newton algorithm.

When $S \approx \operatorname{sign}(S)$, the quantity $\|(\frac{S+S^{-1}}{2})A - A(\frac{S+S^{-1}}{2})\|$ makes a good estimate of the right-hand side of (21). The bound (22) is easily computed or estimated from the known values of A and S. However, these expressions are prone to subtractive cancellation of significant digits.

The quantity $\|E_{21}\|$ is related by (18) to perturbations in the stable invariant subspace. The bounds (21) and (22) are a fortiori bounds on $\|E_{21}\|$, but, as the $(1,2)$ block of (25) suggests, they tend to be pessimistic overestimates of $\|E_{21}\|$ if $\|S\| \gg 1$.

6. The Newton iteration for the computation of the matrix sign function. There are several numerical methods for computing the matrix sign function [2, 25]. Among the simplest and most commonly used is the Newton–Raphson method for a root of $X^2 - I$ starting with initial guess $X_0 = A$ [34, 35]. It is easily implemented using matrix inversion subroutines from widely available, high-quality linear algebra packages like LAPACK [1, 2]. It has been extensively studied and many variations have been suggested [2, 4, 5, 9, 18, 27, 25, 26, 28].

ALGORITHM 1 (*Newton iteration* (*without scaling*)).

$X_0 = A$
FOR $k = 0, 1, 2, \ldots$
$\quad X_{k+1} = (X_k + X_k^{-1})/2$

If A has no eigenvalues on the imaginary axis, then Algorithm 1 converges globally and locally quadratically in a neighborhood of $\operatorname{sign}(A)$ [28]. Although the iteration ultimately converges rapidly, initially convergence may be slow. However, the initial

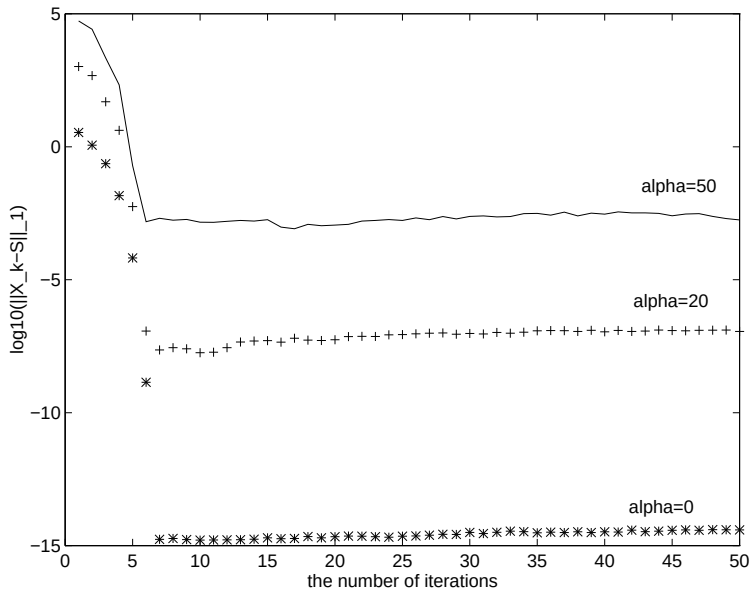


FIG. 1. $\|X_k - \operatorname{sign}(A)\|_1$ *in Example* 2.

convergence rate (and numerical stability) may be improved by scaling [2, 5, 9, 18, 27, 25, 26, 28]. A common choice is to scale X_k $1/|\det(X_k)|^{(1/n)}$ [9].

Theorem 3.2 shows that the first-order perturbation of $\operatorname{sign}(A)$ may be as large as $\|\operatorname{sign}(A)\|^2\epsilon$, where ϵ is the relative uncertainty in A. (If there is no other uncertainty, then ϵ is at least as large as the round-off unit of the finite precision arithmetic.) Thus, it is reasonable to stop the Newton iteration when

$$\|X_{k+1} - X_k\|_1 \le C\epsilon\|X_{k+1}\|_1^2. \tag{26}$$

The ad hoc constant C is chosen to avoid extreme situations, e.g., $C = 1000n$. This choice of C works well in our numerical experiments up to $n = 700$. Experience shows, furthermore, that it is often advantageous to take an extra step of the iteration after the stopping criterion is satisfied.

Example 2. This example demonstrates our stopping criterion. Algorithm 1 was implemented in MATLAB 4.1 on an HP 715/33 workstation with floating point relative accuracy `eps` $= 2 \times 10^{-16}$. We constructed a 10×10 matrix $A = QRQ^H$, where Q is a random unitary matrix and R an upper triangular matrix with diagonal elements $-1 \pm 0.2i, -2.0, -2.5, -3.0, -4.0, -4.5, 2 \pm 0.2i, 6.0$, a parameter α in the $(k, k+2)$ position and zero everywhere else. We chose α such that the norm $\|\operatorname{sign}(A)\|_1$ varies from small to large.

The typical behavior of the error $\|X_k - \operatorname{sign}(A)\|$ is that it goes down and then becomes stationary. This behavior is shown in Figure 1 for the cases $\alpha = 0$, $\alpha = 20$, and $\alpha = 50$ in which $\|\operatorname{sign}(A)\|_1$ is 3×10^0, 2×10^3, and 6×10^4, respectively.

Stopping criterion (26) is satisfied with $C = 1000n$ at the 8th step for $\alpha = 0$ and at the 7th step for $\alpha = 20$ and $\alpha = 50$. Taking one extra step would stop at the 9th step for $\alpha = 0$ and at the 8th step for $\alpha = 20$ and $\alpha = 50$.

In exact arithmetic, the stable and unstable invariant subspaces of the iterates X_k are the same as those of A. However, in finite precision arithmetic, rounding

errors perturb these subspaces. The numerical stability of the Newton iteration for computing the stable invariant subspace was analyzed in [8]; we give an improved error bound here.

Let X and X^+ be, respectively, the computed kth and $(k+1)$st iterates of the Newton iteration starting from

$$X_0 = A = Q \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} Q^H.$$

Suppose that X and X^+ have the form

$$X = Q \begin{bmatrix} X_{11} & X_{12} \\ E_{21} & X_{22} \end{bmatrix} Q^H, \qquad X^+ = Q \begin{bmatrix} X_{11}^+ & X_{12}^+ \\ E_{21}^+ & X_{22}^+ \end{bmatrix} Q^H. \tag{27}$$

A successful rounding error analysis must establish the relationship between E_{21}^+ and E_{21}. To do so we assume that some stable algorithm is applied to compute the inverse X^{-1} in the Newton iteration. More precisely, we assume that X^+ satisfies

$$X^+ = \frac{(X + E_X) + (X + E_X)^{-1}}{2} + E_Z, \tag{28}$$

where

$$\|E_X\| \le c\epsilon\|X\|, \tag{29}$$

$$\|E_Z\| \le c\epsilon(\|X\| + \|X^{-1}\|) \tag{30}$$

for some constant c. Note that this is a nontrivial assumption. Ordinarily, if Gaussian elimination with partial pivoting is used to compute the inverse, the above error bound can be shown to hold only for each column separately [8, 38]. The better inversion algorithms applied to "typical" matrices satisfy this assumption [38, p. 151], but it is difficult to determine if this is always the case [31, pp. 22–26], [20, p. 150].

Write E_X and E_Z as

$$E_X = Q \begin{bmatrix} E'_{11} & E'_{12} \\ E'_{21} & E'_{22} \end{bmatrix} Q^H, \tag{31}$$

$$E_Z = Q \begin{bmatrix} E''_{11} & E''_{12} \\ E''_{21} & E''_{22} \end{bmatrix} Q^H. \tag{32}$$

The following theorem bounds $\|E_{21}\|$ and indirectly the perturbation in the stable invariant subspace.

THEOREM 6.1. *Let X, X^+, E_X, and E_Z be as in* (27), (28), (31), *and* (32). *If* $\frac{1}{2} < 1 - c\epsilon\|X\|\|X_{11}^{-1}\|$, $\frac{1}{2} < 1 - c\epsilon\|X\|\|X_{22}^{-1}\|$, *and*

$$0 < \eta = 1 - 4(\|E_{21}\| + c\epsilon\|X\|)\|X_{22}^{-1}\|\|X_{11}^{-1}\|(\|X_{12}\| + c\epsilon\|X\|),$$

where c is as in (29) *and* (30), *then*

$$\|E_{21}^+\| \le \frac{1}{2}(\|E_{21}\| + c\epsilon\|X\|)\left(1 + \frac{4\|X_{22}^{-1}\|\|X_{11}^{-1}\|}{\eta}\right) + c\epsilon(\|X\| + \|X^{-1}\|).$$

Proof. We start with (28). In fact, the relationship between E_{21} and E_{21}^+ follows from applying the explicit formula for the inverse of $(X + E_X)$ in [32]:

$$Q^H(X+E_X)^{-1}Q = \begin{bmatrix} \tilde{X}_{11}^{-1} + \tilde{X}_{11}^{-1}\tilde{X}_{12}\tilde{X}_c^{-1}(E_{21}+E'_{21})\tilde{X}_{11}^{-1} & -\tilde{X}_{11}^{-1}\tilde{X}_{12}\tilde{X}_c^{-1} \\ -\tilde{X}_c^{-1}(E_{21}+E'_{21})\tilde{X}_{11}^{-1} & \tilde{X}_c^{-1} \end{bmatrix}.$$

Here,

$$\begin{aligned} \tilde{X}_{11} &= X_{11} + E'_{11}, \\ \tilde{X}_{12} &= X_{12} + E'_{12}, \\ \tilde{X}_{22} &= X_{22} + E'_{22}, \\ \tilde{X}_c &= \tilde{X}_{22} - (E_{21}+E'_{21})\tilde{X}_{11}^{-1}\tilde{X}_{12}. \end{aligned}$$

Then

$$\begin{aligned} X_{11}^+ &= \frac{1}{2}(\tilde{X}_{11} + \tilde{X}_{11}^{-1} + \tilde{X}_{11}^{-1}\tilde{X}_{12}\tilde{X}_c^{-1}(E_{21}+E'_{21})\tilde{X}_{11}^{-1}) + E''_{11}, \\ X_{12}^+ &= \frac{1}{2}(\tilde{X}_{12} - \tilde{X}_{11}^{-1}\tilde{X}_{12}\tilde{X}_c^{-1}) + E''_{12}, \\ E_{21}^+ &= \frac{1}{2}((E_{21}+E'_{21}) - \tilde{X}_c^{-1}(E_{21}+E'_{21})\tilde{X}_{11}^{-1}) + E''_{21}, \\ X_{22}^+ &= \frac{1}{2}(\tilde{X}_{22} + \tilde{X}_c^{-1}) + E''_{22}. \end{aligned} \tag{33}$$

Using the Neumann lemma that if $\|B\| < 1$, then $\|(I-B)^{-1}\| < (1-\|B\|)^{-1}$ [17], we have

$$\|\tilde{X}_{11}^{-1}\| \le \frac{\|X_{11}^{-1}\|}{1 - \|X_{11}^{-1}\|\|E'_{11}\|} \le \frac{\|X_{11}^{-1}\|}{1 - c\epsilon\|X_{11}^{-1}\|\|X\|} \le 2\|X_{11}^{-1}\|.$$

The following inequalities are established similarly:

$$\begin{aligned} &\|\tilde{X}_{22}^{-1}\| \le 2\|X_{22}^{-1}\|, \\ &\|\tilde{X}_{12}\| \le \|X_{12}\| + c\epsilon\|X\|, \\ &\|\tilde{X}_c^{-1}\| \le \frac{\|\tilde{X}_{22}^{-1}\|}{1 - \|\tilde{X}_{22}^{-1}\|(\|E_{21}\| + \|E'_{21}\|)\|\tilde{X}_{11}^{-1}\|\|\tilde{X}_{12}\|} \le \frac{2\|X_{22}^{-1}\|}{\eta}. \end{aligned}$$

Inserting these inequalities into (33), we obtain

$$\|E_{21}^+\| \le \frac{1}{2}(\|E_{21}\| + c\epsilon\|X\|)\left(1 + \frac{4\|X_{22}^{-1}\|\|X_{11}^{-1}\|}{\eta}\right) + \|E''_{21}\|. \qquad \square$$

The bound in Theorem 6.1 is stronger than the bound of Byers in [8]. It follows from (19) and Theorem 6.1 that if

$$\frac{\|E_{21}^+\|}{\operatorname{sep}(X_{11}^+, X_{22}^+)} \le \frac{\|E_{21}\|}{\operatorname{sep}(X_{11}, X_{22})},$$

then rounding errors in a step of Newton corrupt the stable invariant subspace by no more than one might expect from the perturbation E_{21} in (27). The term $\operatorname{sep}(X_{11}^+, X_{22}^+)$ is dominated by

$$\operatorname{sep}\left(\frac{X_{11} + X_{11}^{-1}}{2}, \frac{X_{22} + X_{22}^{-1}}{2}\right).$$

So to guarantee that rounding errors in the Newton iteration do little damage, the factors in the bound of Theorem 6.1, $\|X_{11}^{-1}\|\|X_{22}^{-1}\|$ and $(\|X\| + \|X^{-1}\|)$, should be small enough so that

$$\|E_{12}^{+}\| \leq \frac{\operatorname{sep}\left(\frac{X_{11}+X_{11}^{-1}}{2}, \frac{X_{22}+X_{22}^{-1}}{2}\right)}{\operatorname{sep}(X_{11}, X_{22})}\|E_{21}\|. \tag{34}$$

Very roughly speaking, to have numerical stability throughout the algorithm, neither $\|X_{11}^{-1}\|\|X_{22}^{-1}\|$ nor $(\|X\| + \|X^{-1}\|)$ should be much larger than $1/\operatorname{sep}(A_{11}, A_{22})$.

The following example from [4] demonstrates numerical instability that can be traced to the violation of inequality (34).

Example 3. Let

$$A_{11} = \begin{bmatrix} 1-\alpha & & & \alpha \\ \alpha & 1-\alpha & & \\ & \ddots & \ddots & \\ & & \alpha & 1-\alpha \end{bmatrix}$$

be a 10×10 real matrix, and let $A_{22} = -A_{11}^{T}$. Form $R = \left[\begin{smallmatrix} A_{11} & A_{12} \\ E_{21} & A_{22} \end{smallmatrix}\right]$ and $A = QRQ^{T}$, where the orthogonal matrix Q is chosen to be the unitary factor of the QR factorization of a matrix with entries chosen randomly uniformly distributed in the interval $[0, 1]$. The parameter α is taken as $\alpha = (1 - 10^{-5})/2$ so that there are two eigenvalues of A close to the imaginary axis from the left and right sides. The entries of A_{12} are chosen randomly uniformly distributed in the interval $[0, 1]$ too. The entries of E_{21} are chosen randomly uniformly distributed in the interval $[0, \text{eps}]$, where $\text{eps} = 2 \times 10^{-16}$ is the machine precision.

In this example, $\operatorname{sep}(A_{11}, A_{22}) = 2.0 \times 10^{-5}$ and $\sigma_{min}(A) = 3.4 \times 10^{-10}$. Table 1 shows the evolution of $\|E_{21}\|_1/\operatorname{sep}(X_{11}, X_{22})$ during the Newton iteration starting with $X_0 = A$ and $X_0 = R$, respectively, where E_{21} is as in (27). The norm is taken to be the 1-norm. Because $\|A_{11}^{-1}\|_1\|A_{22}^{-1}\|_1 = 1.0 \times 10^{10}$, $\|A^{-1}\|_1 = 2.3 \times 10^{9}$, inequality (34) is violated in the first step of the Newton iteration for the starting matrix A, which is shown in the first column of the table. Newton's method never recovers from this.

It is remarkable, however, that Newton's method applied to R directly seems to recover from the loss in accuracy in the first step. The second column shows that although $\|E_{21}\|_1/\operatorname{sep}(X_{11}, X_{22}) = 1.6 \times 10^{-7}$ at the first step, it is reduced by the factor $1/2$ every step until it reaches 1.7×10^{-12}, which is approximately $\|E_{21}\|_1/\operatorname{sep}(A_{11}, A_{22})$. Observe that in this case the perturbation E_{21}'' in E_Z as in (28) is zero and $\|E_{21}^{+}\|_1$ is dominated by $\frac{1}{2}(\|E_{21}\|_1 + \|X_{22}^{-1}E_{21}X_{11}^{-1}\|_1)$. It is surprising to see that from the second step on $\|X_{11}^{-1}E_{21}X_{22}^{-1}\|_1$ is as small as eps, since A_{11}^{-1} and A_{22}^{-1} do not explicitly appear in the term $X_{11}^{-1}E_{21}X_{22}^{-1}$ after the first step.

Our analysis suggests that the Newton iteration may be unstable when X_k is ill conditioned. To overcome this difficulty the Newton iteration may be carried out with a shift along the imaginary line. In this case we must use complex arithmetic.

ALGORITHM 2 (Newton iteration with shift).

$X_0 = A - \beta i I$

FOR $k = 0, 1, 2, \ldots$

$\quad X_{k+1} = (X_k + X_k^{-1})/2$

END

TABLE 1
Evolution of $\|E_{21}\|_1/\operatorname{sep}(X_{11}, X_{22})$ *in Example* 3.

k	$\|E_{21}\|_1/\operatorname{sep}(X_{11}, X_{22})$		$\operatorname{sep}(X_{11}, X_{22})$
	A	R	
0	8.7451e−11	7.0512e−11	2.0000e−05
1	7.7083e−07	1.5779e−07	1.0955e+00
2	5.0378e−07	1.0905e−07	7.9263e−01
3	1.2093e−07	2.5501e−08	1.6948e+00
4	8.3733e−08	1.2150e−08	1.7786e+00
5	7.3034e−08	5.4025e−09	2.0000
6	7.3164e−08	2.7012e−09	2.0000
7	7.2020e−08	1.3506e−09	2.0000
8	7.1731e−08	6.7532e−10	2.0000
9	7.1866e−08	3.3766e−10	2.0000
10	7.1888e−08	1.6883e−10	2.0000
11	7.1909e−08	8.4426e−11	2.0000
12	7.1926e−08	4.2231e−11	2.0000
13	7.1934e−08	2.1151e−11	2.0000
14	7.1938e−08	1.0646e−11	2.0000
15	7.1938e−08	5.4637e−12	2.0000
16	7.1937e−08	3.0055e−12	2.0000
17	7.1938e−08	2.0001e−12	2.0000
18	7.1937e−08	1.7474e−12	2.0000
19	7.1937e−08	1.7291e−12	2.0000
20	7.1937e−08	1.7290e−12	2.0000
21	7.1937e−08	1.7290e−12	2.0000

The real parameter β is chosen such that $\sigma_{min}(A-\beta iI)$ is not small. For Example 2, when β is taken to be 0.8, we have $\|E_{21}\|_1/\operatorname{sep}(X_{11}, X_{22}) = 7.3 \times 10^{-12}$ for $k = 21$. Then by our analysis the computed invariant subspace is guaranteed to be accurate.

7. Conclusions. We have given a first-order perturbation theory for the matrix sign function and an error analysis for Newton's method to compute it. This analysis suggests that computing the stable (or unstable) invariant subspace of a matrix with the Newton iteration in most circumstances yields results as good as those obtained from the Schur form.

Acknowledgments. The authors would like to express their thanks to N. Higham for valuable comments on an earlier draft of the paper and Z. Bai and P. Benner for helpful discussions.

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] Z. BAI AND J.W. DEMMEL, *Design of a parallel nonsymmetric eigenroutine toolbox, Part* I, in Proc. of the Sixth SIAM Conference on Parallel Processing for Scientific Computing, R. F. Sincovec, ed., SIAM, Philadelphia, PA, 1993. Long version available as Technical report CSD-92-718, Dept. of Comp. Sci., Univ. of California, Berkeley, CA.

[3] Z. BAI AND J.W. DEMMEL, *Design of a Parallel Nonsymmetric Eigenroutine Toolbox, Part* II, Technical report 95-11, Dept. of Mathematics, University of California, Berkeley, CA, 1995.

[4] Z. BAI, J.W. DEMMEL, AND M. GU, *Inverse free parallel spectral divide and conquer algorithms for nonsymmetric eigenproblems*, Numer. Math., to appear.

[5] L.A. BALZER, *Accelerated convergence of the matrix sign function method of solving Lyapunov, Riccati and other matrix equations*, Internat. J. Control, 32 (1980), pp. 1057–1078.

[6] A.N. Beavers and E.D. Denman, *A computational method for eigenvalues and eigenvectors of a matrix with real eigenvalues*, Numer. Math., 21 (1973), pp. 389–396.
[7] S. Boyd and V. Balakrishnan, *A regularity result and a quadratically convergent algorithm for computing its L_∞ norm*, Systems Control Lett., 15 (1990), pp. 1–7.
[8] R. Byers, *Numerical stability and instability in matrix sign function based algorithms*, in Computational and Combinatorial Methods in Systems Theory, C.I. Byrnes and A. Lindquist, eds., Elsevier, North–Holland, 1986, pp. 185–199.
[9] R. Byers, *Solving the algebraic Riccati equation with the matrix sign function*, Linear Algebra Appl., 85 (1987), pp. 267–279.
[10] R. Byers, *A bisection method for measuring the distance of a stable matrix to the unstable matrices*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 875–881.
[11] J.P. Charlier and P. Van Dooren, *A systolic algorithm for Riccati and Lyapunov equations*, Math. Control Signals Systems, 2 (1989), pp. 109–136.
[12] B. De-Sz. Nagy, *Perturbations des transformations autoadjointes dans l'espace de Hilbert*, Comment. Math. Helv., 19 (1947), pp. 347–366.
[13] E.D. Denman and A.N. Beavers, *The matrix sign function and computations in systems*, Appl. Math. Comput., 2 (1976), pp. 63–94.
[14] E.D. Denman and J. Leyva-Ramos, *Spectral decomposition of a matrix using the generalized sign matrix*, Appl. Math. Comput., 8 (1981), pp. 237–250.
[15] 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.
[16] J.D. Gardiner and A.J. Laub, *Parallel algorithms for algebraic Riccati equations*, Internat. J. Control, 54 (1991), pp. 1317–1333.
[17] G.H. Golub and C.F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[18] N.J. Higham, *Computing the polar decomposition—with applications*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 1160–1174.
[19] N.J. Higham, *Newton's method for the matrix square root*, Math. Comp., 46 (1986), pp. 537–549.
[20] V. Kahan, *A survey of error analysis*, in Proc. of the IFIP Congress, North–Holland, Amsterdam, 1971, pp. 1214–1239.
[21] T. Kato, *On the convergence of the perturbation method*, I, Progr. Theory Phys., 4 (1949), pp. 514–523.
[22] T. Kato, *On the convergence of the perturbation method*, II, Progr. Theory Phys., 5 (1950), pp. 207–212.
[23] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1980.
[24] C. Kenney and A.J. Laub, *Polar decompositions and matrix sign function condition estimates*, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 488–504.
[25] C. Kenney and A.J. Laub, *Rational iterative methods for the matrix sign function*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 273–291.
[26] C. Kenney and A.J. Laub, *Matrix-sign algorithms for Riccati equations*, IMA J. Math. Control Inf., 9 (1992), pp. 331–344.
[27] C. Kenney and A.J. Laub, *On scaling Newton's method for polar decompositions and the matrix sign function*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 688–706.
[28] C. Kenney and A.J. Laub, *The matrix sign function*, IEEE Trans. Automat. Control, 40 (1995), pp. 1330–1348.
[29] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, Academic Press, Orlando, FL, 1985.
[30] R. Mathias, *Condition estimation for the matrix sign function via the Schur decomposition*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 565–578.
[31] W. Miller and C. Wrathall, *Software for Roundoff Analysis of Matrix Algorithms*, Academic Press, New York, 1980.
[32] D.V. Ouellette, *Schur complement and statistics*, Linear Algebra Appl., 36 (1981), pp. 187–295.
[33] P. Pandey, C. Kenney, and A.J. Laub, *A parallel algorithm for the matrix sign function*, Internat. J. High Speed Comput., 2 (1990), pp. 181–191.
[34] J.D. Roberts, *Linear Model Reduction and Solution of Algebraic Riccati Equation by Use of the Sign Function*, Technical report CUED/B-Control, TR-13, Cambridge University, Cambridge, UK, 1971.
[35] J.D. Roberts, *Linear model reduction and solution of the algebraic Riccati equation by use of the sign function*, Internat. J. Control, 32 (1980), pp. 677–687.
[36] M. Rosenblum, *On the operator equation "$BX - XA = Q$,"* Duke Math. J., 23 (1956), pp.

263–269.
[37] G.W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.
[38] G.W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.
[39] G.W. Stewart and Ji-Guang Sun, *Matrix Perturbation Theory*, Academic Press, London, UK, 1990.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 3, pp. 633–652, July 1997

007

IMPLICITLY RESTARTED KRYLOV SUBSPACE METHODS FOR STABLE PARTIAL REALIZATIONS*

IMAD M. JAIMOUKHA† AND EBRAHIM M. KASENALLY‡

Abstract. This paper considers an implicitly restarted Krylov subspace method that approximates a stable, linear transfer function $f(s)$ of order n by one of order m, where $n \gg m$. It is well known that oblique projections onto a Krylov subspace may generate unstable partial realizations. To remedy this situation, the oblique projectors obtained via classical Krylov subspace methods are supplemented with further projectors which enable the formation of stable partial realizations directly from $f(s)$. A key feature of this process is that it may be incorporated into an implicit restart scheme. A second difficulty arises from the fact that Krylov subspace methods often generate partial realizations that contain nonessential modes. To this end, balanced truncation may be employed to discard the unwanted part of the reduced-order model. This paper proposes oblique projection methods for large-scale model reduction that simultaneously compute stable reduced-order models while discarding all nonessential modes. It is shown that both of these tasks may be effected by a single oblique projection process. Furthermore, the process is shown to naturally fit into an implicit restart framework. The theoretical properties of these methods are thoroughly investigated, and exact low-dimensional expressions for the $\mathcal{L}^\infty$-norm of the residual errors are derived. Finally, the behavior of the algorithm is illustrated on two large-scale examples.

Key words. Arnoldi, model reduction, Krylov subspace methods, large-scale systems, implicit restarts

AMS subject classifications. 65F10, 65F15, 93A15, 93B05, 93B07, 93B20

PII. S0895479895279873

1. Introduction.

Model reduction has been practiced widely by engineers, and, until recently, the process was often based on intuition and a sound understanding of the physical principles associated with the modeling task. Chemical engineers assume that mixing is instantaneous and that packed distillation columns may be modeled using discrete trays. Electrical engineers represent transmission lines and eddy currents in the rotor cage of an induction motor by lumped circuits. Mechanical engineers routinely remove the high-frequency vibration modes from models of aircraft wings, turbine shafts, and flexible structures. The purpose of the present paper is to systematize the model reduction of large-scale dynamical systems without any a priori knowledge of their characteristics and to provide computable expressions for the errors incurred during the approximation process.

Consider the dynamical system described by a stable, linear, time-invariant state-space model of the form

$$\dot{x}(t) = Ax(t) + bu(t), \qquad y(t) = cx(t), \tag{1}$$

in which $x(t)$ is the state vector of dimension n and $u(t)$ and $y(t)$ are scalar functions representing the input and the output of the system, respectively. The matrix A and

* Received by the editors January 13, 1995; accepted for publication (in revised form) by P. Van Dooren July 12, 1996.

http://www.siam.org/journals/simax/18-3/27987.html

† Interdisciplinary Research Centre for Process Systems Engineering and Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, London SW7-2BY, U.K. (jaimouka@ps.ic.ac.uk).

‡ Bank of America NT & SA, Exposure Management, EMEA 5209, 26 Elmfield Road, Bromley, Kent BR1-1WA, U.K. The research contained in this paper was performed while this author was with the Interdisciplinary Research Centre for Process Systems Engineering, Imperial College of Science, Technology and Medicine, London SW7-2BY, U.K. (kasenall@ps.ic.ac.uk).

vectors b and c are real, with their dimensions fixed by those of $x(t)$, $u(t)$, and $y(t)$. It should be noted that the matrix A will be large and sparse in the following. Applying Laplace transforms to the system in (1) for zero initial conditions leads to a transfer function given by $f(s) = c(sI - A)^{-1}b$. The task of any model-reduction algorithm is then to find a stable approximate model of the form

$$\dot{x}_m(t) = A_m x_m(t) + b_m u(t), \qquad y_m(t) = c_m x_m(t), \tag{2}$$

in which $x_m(t) \in \mathbb{R}^m$ and $m \ll n$ and where the associated low-order transfer function is given by $f_m(s) = c_m(sI - A_m)^{-1}b_m$. Well-established model reduction methods such as the optimal Hankel norm [8] and the balanced truncation [20] begin by solving the linear matrix equations

$$AP + PA^T + bb^T = 0 \qquad \text{and} \qquad A^TQ + QA + c^Tc = 0, \tag{3}$$

which admit unique symmetric solutions if and only if $\lambda_i(A) + \bar{\lambda}_j(A) \neq 0$ for all i, j, where λ_i denotes the ith eigenvalue and the overbar represents the complex conjugate. Approximating large-scale systems via such methods is intractable because of the prohibitive storage requirements and a computational burden of $O(n^3)$ for each Lyapunov equation. The approach developed here has its roots in determining low-rank approximate solutions to (3) via the application of classical Krylov subspace methods [1, 18].

Recently, several schemes using Krylov subspace methods have been proposed for the task of large-scale model reduction. The objective of such algorithms has been to replace a high-order model by an mth-order low-dimensional approximation while effecting $O(m^2n)$ operations. For example, in [1], this type of model reduction formed an integral part of control system synthesis where high-order controllers had to be avoided because of certain engineering considerations. Krylov subspace model reduction schemes have recently been employed in linear circuit analysis, where they allow for the accurate and efficient simulation of large-scale circuits [6]. In our previous work [16, 17, 18], the focus was placed on calculating low-rank approximate solutions to (3). It was shown that these low-rank solutions were the exact controllability and observability Gramians of a dynamical system obtained by perturbing the system in (1). Finally, it was demonstrated that the reduced-order models could be readily computed via an oblique projection process based on the data generated in the course of the iterative process. It is well known, however, that partial realizations computed in this way may be unstable even if $f(s)$ is stable. Thus, one of the aims of this paper is to propose a method that yields a stable reduced-order model by effecting a second oblique projection process which we shall call a stable projection. An interesting feature of this method is that by combining the oblique projectors generated by the iterative process with those from the stable projection, one may compute stable reduced-order models by effecting a single oblique projection on $f(s)$. Consequently, one may cast the algorithm into an implicit restart setting which is similar in philosophy to that proposed in [14, 15].

Stable partial realizations via implicitly restarted Krylov subspace methods have been proposed in [14] in the context of control and in [15] in the setting of linear circuit analysis. The difference between the methods proposed here and those of [14, 15] is that, in the present setting, the integrity of both Arnoldi equations is preserved, whereas [14, 15] focus on preserving only one of these equations. One may not wish to approximate the state transition matrix independently of b and c, such as in a modal reduction scheme, but rather obtain a reduced-order projection process that

uses all the problem data. One advantage of the present approach is that it offers computable expressions for the $\mathcal{L}^\infty$-norm of the residual errors. Another advantage is that the theoretical properties expounded in [18] remain in force irrespective of the number of restarts performed; furthermore, the reduced-order model satisfies a moment-matching property. While the implicit restart strategy presented here focuses on model reduction, this approach may also be used to implicitly restart Krylov subspace schemes for the solution of large-scale Lyapunov equations [16, 17] and approximations to the exponential operator [22].

Krylov subspace methods are known to generate partial realizations which capture the outermost part of the spectrum of A. While this feature may be desirable in some settings, such as for the solution of stiff ordinary differential equations, it often results in reduced-order models that are unable to approximate the low-frequency characteristics of $f(s)$. A second objective addressed in this paper is to extract the nonessential modes (i.e., those modes associated with the outermost part of A's spectrum) from the reduced-order model by effecting a balanced truncation of $f_m(s)$. We make use of a numerically robust variant of Moore's algorithm [20] which does not require the formation of the balancing transformations, known as the square root algorithm [24]. A key feature of the square root model-reduction scheme is that it may be cast as an oblique projection process and when combined with the oblique projectors derived from the iterative process yields projectors that compute a reduced-order approximation to $f(s)$ which retains the essential characteristics of $f_m(s)$. A second feature that will be exploited is that this process may be cast naturally into an implicit restart framework. It is interesting to observe that one may combine the oblique projectors from the iterative process with those obtained from the stable projection and balanced truncation to produce an implicitly restarted Krylov subspace method that computes stable partial realizations that accurately replicate the low-frequency characteristics of $f(s)$.

Related to this work are those of [4, 5], which use Krylov subspace methods to obtain bases for the controllability and observability spaces; furthermore, in [5] Boley and Golub presented a means of computing a minimal realization of a linear dynamical system from the coefficients generated in the course of the Lanczos process. The Lanczos process was also exploited by Parlett in [21] to obtain minimal realizations. In that paper, the rank of the Hankel matrix was used to determine the order of the minimal realization; furthermore, it was demonstrated that a minimal realization could be constructed from the data generated by the Lanczos process. A similar approach was adopted in [12], in which the minimal realization and its order were found to be related to the different types of breakdown encountered in the Lanczos process. Here too the onset of breakdown was given in terms of properties of the Hankel matrix. Recently, the presentations in [3, 26] review the use of projection methods for large-scale control problems. Each paper suggests the use of Krylov subspace methods as an effective tool for the model reduction of large-scale linear dynamical systems; however, no algorithms were provided.

The following summarizes the structure of this paper. Section 2 briefly describes the application of classical Krylov subspace techniques to the model-reduction problem. Section 3 considers a general implicit restart structure for arbitrary transformations, a theoretical analysis is presented and computable $\mathcal{L}^\infty$-norms of the error expressions are derived. Section 4 uses the framework of section 3 to produce several implicitly restarted Krylov subspace algorithms based on particular transformations such as stable projections or balancing transformations. Two numerical experiments

expounding the benefits of an implicitly restarted Krylov subspace method are found in section 5, and section 6 contains the conclusions.

2. Oblique projection methods for large-scale model reduction. This section considers oblique projection methods onto the m-dimensional Krylov subspaces

$$\mathcal{K}_m(A,b) := \text{span}\left\{[\, b \; Ab \; \cdots \; A^{m-1}b \,]\right\}$$
$$\mathcal{L}_m(A^T,c^T) := \text{span}\left\{[\, c^T \; A^Tc^T \; \cdots \; (A^{m-1})^Tc^T \,]\right\},$$

which are parts of the controllability and observability subspaces, respectively. The class of iterative techniques focused on hinge on the efficient formation of well-conditioned bases for $\mathcal{K}_m(A,b)$ and $\mathcal{L}_m(A^T,c^T)$. To this end, one exploits a modified Gram–Schmidt process known as the Arnoldi process to calculate matrices $V_m = [v_1, v_2, \ldots, v_m]$ and $W_m = [w_1, w_2, \ldots, w_m]$, whose columns form an orthogonal basis for each of the Krylov subspaces as well as unit vectors v_{m+1} and w_{m+1}, which are orthogonal to V_m and W_m, respectively [2]. An outline of the Arnoldi process and its application to several large-scale numerical linear algebra problems may be found in [4, 7, 16, 17, 18, 19, 23]. By construction, the Arnoldi process associated with $\mathcal{K}_m(A,b)$ produces an $m \times m$ upper Hessenberg matrix H_m; furthermore, one readily verifies that b and A satisfy

$$b = V_m l_m, \tag{4}$$
$$AV_m = V_m H_m + \widetilde{V}_m \widetilde{H}_m, \tag{5}$$

in which $l_m = e_1\|b\|_2$, $\widetilde{V}_m = v_{m+1}$, and $\widetilde{H}_m = h_{m+1,m}e_m^T$ and where $h_{m+1,m}$ is a nonnegative scalar and e_1 and e_m are, respectively, the first and last columns of the m-dimensional identity matrix. From (5), it is easy to see that $H_m = V_m^TAV_m$ since $[V_m \; \widetilde{V}_m]$ is part of an orthogonal matrix. In what follows, we refer to (4) and (5) as the controllability Arnoldi equations. Similarly, associated with $\mathcal{L}_m(A^T,c^T)$, the Arnoldi process generates a lower Hessenberg matrix G_m; in this setting, c and A then satisfy

$$c^T = W_m k_m^T, \tag{6}$$
$$A^TW_m = W_m G_m^T + \widetilde{W}_m \widetilde{G}_m^T, \tag{7}$$

in which $k_m = \|c\|_2 e_1^T$, $\widetilde{W}_m = w_{m+1}$, and $\widetilde{G}_m^T = g_{m,m+1}e_m^T$ and where $g_{m,m+1}$ is a nonnegative scalar. It follows that $G_m = W_m^TAW_m$ since $[W_m \; \widetilde{W}_m]$ is part of an orthogonal matrix. In the development below, we refer to (6) and (7) as the observability Arnoldi equations. A key difference between the restart process proposed in this paper and those of [14, 15] is that the present approach aims to preserve the integrity of (4)–(7). The main advantage of this approach is that it allows simple residual error expressions to be computed. In contrast, [14, 15] focus on preserving (5) and (7), and no residual error expressions are provided. Furthermore, we show that the implicitly restarted approximate model continues to satisfy a (modified) moment-matching property enjoyed by the oblique projection approximation without restarts.

2.1. Model reduction using Krylov subspace methods. The aim of this section is to consider the Krylov subspace techniques described above to provide

computationally efficient model-reduction schemes for large-scale dynamical systems. Denoting the transfer function corresponding to the model (1) by $f(s)$, then

$$f(s) = c(sI - A)^{-1}b \stackrel{s}{=} \left[\begin{array}{c|c} A & b \\ \hline c & 0 \end{array}\right], \quad A \in \mathbb{R}^{n\times n}, \quad b, c^T \in \mathbb{R}^n. \tag{8}$$

The model-reduction task determines a reduced-order model given by

$$f_m(s) = c_m(sI - A_m)^{-1}b_m \stackrel{s}{=} \left[\begin{array}{c|c} A_m & b_m \\ \hline c_m & 0 \end{array}\right], \quad A_m \in \mathbb{R}^{m\times m}, \quad b_m, c_m^T \in \mathbb{R}^m,$$

which approximates the high-dimensional model $f(s)$, where $m \ll n$.

Rewriting (8) as $f(s) = cf_b(s) = f_c(s)b$, where $f_b(s) = (sI - A)^{-1}b$ and $f_c(s) = c(sI - A)^{-1}$, permits us to consider $f_b(s)$ and $f_c(s)$ as the solutions to the coupled linear systems

$$(sI - A)f_b(s) = b \qquad \text{and} \qquad f_c(s)(sI - A) = c, \tag{9}$$

respectively. The focus of what follows is to approximate $f(s)$ by obtaining approximate solutions $f_{b,m}(s)$ and $f_{c,m}(s)$ to the linear systems (9). These approximate solutions are constructed to satisfy the following two conditions. (1) $f_{b,m}(s) \in \mathcal{K}_m(A, b)$, i.e., $f_{b,m}(s) = V_m h_m(s)$, such that $\mathcal{L}_m(A^T, c^T) \perp \{(sI - A)f_{b,m}(s) - b\}$. (2) $f_{c,m}^T(s) \in \mathcal{L}_m(A^T, c^T)$, i.e., $f_{c,m}(s) = g_m(s)W_m^T$, such that $\{f_{c,m}(s)(sI - A) - c\} \perp \mathcal{K}_m(A, b)$. Since $f_{b,m}(s)$ and $f_{c,m}(s)$ are approximate solutions to the linear systems in (9) and $f(s) = cf_b(s) = f_c(s)b$, we will consider $f_{m,1}(s) := cf_{b,m}(s)$ and $f_{m,2}(s) := f_{c,m}(s)b$ as approximations to $f(s)$. The problem we wish to solve may be stated as follows.

PROBLEM 2.1. *Find approximate solutions $f_{b,m}(s) = V_m h_m(s)$ and $f_{c,m}(s) = g_m(s)W_m^T$ to* (9) *which satisfy the Galerkin-type conditions*

$$W_m^T\{(sI - A)V_m h_m(s) - b\} = 0\,\forall s, \qquad \{g_m(s)W_m^T(sI - A) - c\}V_m = 0\,\forall s. \tag{10}$$

For convenience, define the matrices

$$\begin{aligned} \widehat{H}_m &:= T_m^{-1}W_m^T A V_m = H_m + T_m^{-1}W_m^T\widetilde{V}_m\widetilde{H}_m, \\ \widehat{G}_m &:= W_m^T A V_m T_m^{-1} = G_m + \widetilde{G}_m\widetilde{W}_m^T V_m T_m^{-1} \end{aligned}$$

for nonsingular $T_m := W_m^T V_m$ and observe that $\widehat{H}_m$ and $\widehat{G}_m$ are upper and lower Hessenberg, respectively. The following theorem gives the solution to Problem 2.1.

THEOREM 2.1 (see [18]). *Suppose that m steps of the Arnoldi process have been taken and that T_m is nonsingular. Then*

1. *the Galerkin conditions in* (10) *are satisfied if and only if $h_m(s) = (sI - \widehat{H}_m)^{-1}l_m$ and $g_m(s) = k_m(sI - \widehat{G}_m)^{-1}$. Under these conditions, the residual error $\mathcal{L}^\infty$-norms are*

$$\|b - (sI - A)V_m h_m(s)\|_\infty = \left\| \begin{bmatrix} T_m^{-1}W_m^T\widetilde{V}_m \\ 1 \end{bmatrix} \widetilde{H}_m h_m(s) \right\|_\infty, \tag{11}$$

$$\left\|c - g_m(s)W_m^T(sI - A)\right\|_\infty = \left\| g_m(s)\widetilde{G}_m \begin{bmatrix} \widetilde{W}_m^T V_m T_m^{-1} & 1 \end{bmatrix} \right\|_\infty; \tag{12}$$

2. *the approximations*

$$f_{m,1}(s) = cV_m h_m(s) \stackrel{s}{=} \left[\begin{array}{c|c} T_m^{-1}W_m^T A V_m & T_m^{-1}W_m^T b \\ \hline cV_m & 0 \end{array}\right] = \left[\begin{array}{c|c} \widehat{H}_m & l_m \\ \hline k_m T_m & 0 \end{array}\right] \tag{13}$$

and

$$f_{m,2}(s) = g_m(s) W_m^T b \stackrel{s}{=} \left[\begin{array}{c|c} W_m^T A V_m T_m^{-1} & W_m^T b \\ \hline c V_m T_m^{-1} & 0 \end{array}\right] = \left[\begin{array}{c|c} \widehat{G}_m & T_m l_m \\ \hline k_m & 0 \end{array}\right] \tag{14}$$

are different realizations of the same transfer function, namely, $f_m(s) = f_{m,1}(s) = f_{m,2}(s)$ *for all* s.

Remark. 2.1. Throughout this paper, we assume that T_m is nonsingular, which is equivalent to a breakdown-free Arnoldi process. Furthermore, for simplicity, we only consider the model reduction of single-input single-output transfer functions. For the model reduction of multi-input multioutput systems, one needs to resort to block Arnoldi schemes. For more details of such processes, including breakdown, we refer the reader to [4, 18] and the references therein.

Thus, for the mth-order approximate model described by (2), we can take either $f_m(s) = f_{m,1}(s)$ or $f_m(s) = f_{m,2}(s)$, where $f_{m,1}(s)$ and $f_{m,2}(s)$ are given by (13) and (14), respectively. The following procedure summarizes an oblique projection method for model reduction of large-scale systems.

ALGORITHM 2.1 (Krylov subspace model-reduction algorithm).

- *Start: Specify tolerances* $\gamma > 0$ *and* $\epsilon > 0$; *set an integer parameter* m.
- *Perform* m *steps of the Arnoldi process with* (A, b) *to produce* H_m, $\widetilde{H}_m$, V_m, $\widetilde{V}_m$, *and* l_m.
- *Perform* m *steps of the Arnoldi process with* (A^T, c^T) *to produce* G_m, $\widetilde{G}_m$, W_m, $\widetilde{W}_m$, *and* k_m.
- *Form the reduced-order model from either* (13) *or* (14).
- *Test the* $\mathcal{L}^\infty$*-norm of the errors in* (11) *and* (12); *if either* (11) $> \gamma$ *or* (12) $> \epsilon$, *increase* m *and continue the Arnoldi processes.*

The reduced-order models given by (13) and (14) are computed readily from the data generated in the course of the Arnoldi processes. It is known that such models may be unstable even if $f(s)$ is stable; furthermore, such partial realizations often contain modes associated with the outer part of the spectrum of A. The following section presents a framework able to remove such undesirable features by the application of further oblique projectors within an implicit restart setting.

3. A general implicit restart framework. Implicit restart schemes were first proposed to compute a few desired eigenvalues of large sparse nonsymmetric matrices [25]. More recently, they have been exploited to compute stable partial realizations in the setting of control problems and linear circuit analysis [14, 15]. The aim of this section is to propose a general implicit restart framework based on the Arnoldi process. A key feature of the present approach, which differs from existing implicit restart schemes, is that the integrity of the controllability and observability Arnoldi equations (4)–(7) is preserved. In contrast, existing methods focus on preserving (5) and (7), which makes it difficult to establish system theoretic connections between $f_m(s)$ and $f(s)$. In the eigenvalue setting, preserving the integrity of either (4) or (6) is not essential, since the starting vector is an eigenvector estimate and does not form part of the problem data. An advantage offered by preserving the Arnoldi equations is that one may establish computable error expressions that are similar to those of Theorem 2.1; furthermore, one may demonstrate that the implicitly restarted reduced-order model may be obtained by effecting low-rank perturbations to the state-space representation of $f(s)$.

Remark. 3.1. Commonly, implicit restarts refer to restarting the Arnoldi process with updated starting vectors v_1 and w_1. In this work, we take a broader inter-

pretation, so that restarts refer to removing, via oblique projections, all features of $f_m(s)$ that are deemed undesirable in a given application and restarting the (modified) Arnoldi process while preserving the Arnoldi equations (4)–(7).

Suppose that m steps of the Arnoldi process have been taken and that $f_{m,1}(s)$, given in (13), is the reduced-order model obtained upon the application of the oblique projection process in Theorem 2.1. Observe that $f_{m,1}(s)$ is not necessarily stable; furthermore, similar to a power method, the Arnoldi process engenders an $\widehat{H}_m$ whose eigenvalues approximate those of A with large absolute value. The presence of such features makes $f_{m,1}(s)$ unsuitable for many practical applications such as circuit simulation. Furthermore, robust controller design methods based on the small gain theorem require that the actual model (in this case $f(s)$) and the nominal model (in this case $f_{m,1}(s)$) have the same number of poles in the closed right-half complex plane [9, 10]. Since $f(s)$ is stable from the outset, stability of $f_{m,1}(s)$ is required for such methods. Furthermore, if many eigenvalues of $\widehat{H}_m$ approximate those of A with large absolute value, $f_{m,1}(s)$ is a poor reduced-order model since it is unable to replicate the low-frequency characteristics of $f(s)$. Suppose that these undesirable features may be extracted via the application of an additional oblique projection process; namely, define two full-column rank matrices $T_L, T_R \in \mathbb{R}^{m\times r}$ such that $T_L^T T_R = I_r$, where I_r is the $r \times r$ identity matrix and $r < m$. Then the desirable portion of $f_{m,1}(s)$ is given by

$$f_{r,1}(s) \stackrel{s}{=} \left[\begin{array}{c|c} T_L^T \widehat{H}_m T_R & T_L^T l_m \\ \hline k_m T_m T_R & 0 \end{array}\right].$$

The following is referred to as a basis change in the state-space realization of any rational $g(s)$:

$$g(s) \stackrel{s}{=} \left[\begin{array}{c|c} A & b \\ \hline c & d \end{array}\right] \stackrel{T}{\rightarrow} g(s) \stackrel{s}{=} \left[\begin{array}{c|c} TAT^{-1} & Tb \\ \hline cT^{-1} & d \end{array}\right],$$

where T is nonsingular. The next result establishes that the projectors T_L and T_R applied to $f_{m,1}(s)$ may be combined with the oblique projectors generated in the course of the Arnoldi process; furthermore, the composite projectors yield a reduced-order model whose structure is reminiscent of (13).

PROPOSITION 3.1. *Suppose that m steps of the Arnoldi process have been taken and that T_m is nonsingular. Let $T_R = Q_R R_R$ and $(T_m^T)^{-1} T_L = Q_L R_L$ be QR decompositions in which $Q_L, Q_R \in \mathbb{R}^{m\times r}$ are parts of orthogonal matrices and $R_L, R_R \in \mathbb{R}^{r\times r}$ are upper triangular. Define $V_r := V_m Q_R$, $W_r := W_m Q_L$, and $T_r := W_r^T V_r$. Then*

1. *T_r is nonsingular and $T_r^{-1} = (Q_L^T T_m Q_R)^{-1} = R_R R_L^T$,*
2. *$f_{r,1}(s)$ may be expressed as*

$$f_{r,1}(s) \stackrel{s}{=} \left[\begin{array}{c|c} T_r^{-1} W_r^T A V_r & T_r^{-1} W_r^T b \\ \hline cV_r & 0 \end{array}\right]. \tag{15}$$

Proof. Since T_L and T_R have full-column rank, R_L and R_R are nonsingular. Using the QR decompositions and $T_L^T T_R = I_r$, $T_r = Q_L^T T_m Q_R = (R_L^T)^{-1} R_R^{-1}$, from which part 1 follows. By Theorem 2.1, $f_{r,1}(s)$ may be expressed as

$$\begin{aligned} f_{r,1}(s) &\stackrel{s}{=} \left[\begin{array}{c|c} T_L^T T_m^{-1} W_m^T A V_m T_R & T_L^T T_m^{-1} W_m^T b \\ \hline cV_m T_R & 0 \end{array}\right] \\ &\stackrel{s}{=} \left[\begin{array}{c|c} R_R T_L^T T_m^{-1} W_m^T A V_m Q_R & R_R T_L^T T_m^{-1} W_m^T b \\ \hline cV_m Q_R & 0 \end{array}\right] \end{aligned} \tag{16}$$

upon substituting $T_R = Q_R R_R$ and effecting a basis change using the transformation R_R. Substituting $(T_m^T)^{-1} T_L = Q_L R_L$ into (16) and using part 1 readily establishes the claim in part 2. ∎

Remark. 3.2. Alternatively, one may apply the transformations T_L and T_R to the reduced-order model $f_{m,2}(s)$, given in (14), to yield

$$f_{r,2}(s) \stackrel{s}{=} \left[\begin{array}{c|c} T_L^T \widehat{G}_m T_R & T_L^T T_m l_m \\ \hline k_m T_R & 0 \end{array}\right].$$

On effecting the QR decompositions $T_L = Q_L R_L$ and $T_m^{-1} T_R = Q_R R_R$, one readily demonstrates that

$$f_{r,2}(s) \stackrel{s}{=} \left[\begin{array}{c|c} W_r^T A V_r T_r^{-1} & W_r^T b \\ \hline c V_r T_r^{-1} & 0 \end{array}\right], \tag{17}$$

where V_r and W_r are defined in Proposition 3.1.

Observe that V_r and W_r continue to remain orthogonal bases to parts of the controllability and observability subspaces since Q_R and Q_L are parts of orthogonal matrices. From Proposition 3.1, it appears that $f_{r,1}(s)$ may be computed by imposing a Galerkin condition on the residual of (9). Thus, similar to the arguments leading to the statement of Problem 2.1, one defines an approximate solution to the first linear system of (9) of the form $f_{b,r}(s) = V_r h_r(s)$; then we consider $f_{r,1}(s) = c f_{b,r}(s)$ as an approximation to $f(s)$. Similarly, $f_{c,r}(s) = g_r(s) W_r^T$ defines an approximate solution to the second linear system of (9), which leads to an approximation of $f(s)$ given by $f_{r,2}(s) = f_{c,r}(s) b$. The functions $h_r(s)$ and $g_r(s)$ are then computed by imposing Galerkin conditions as shown by the next corollary. For notational convenience, define $\widehat{H}_r := T_r^{-1} W_r^T A V_r$ and $\widehat{G}_r := W_r^T A V_r T_r^{-1}$.

COROLLARY 3.2. *Suppose that the conditions of Proposition* 3.1 *are in force. Then*

1. $h_r(s) = \left(sI - \widehat{H}_r\right)^{-1} T_r^{-1} W_r^T b$ *if and only if* $\{(sI - A) V_r h_r(s) - b\} \perp W_r$.
2. $g_r(s) = c V_r T_r^{-1} \left(sI - \widehat{G}_r\right)^{-1}$ *if and only if* $\left\{c - g_r(s) W_r^T (sI - A)\right\} \perp V_r$.

Proof. By direct calculation,

$$\begin{aligned} W_r^T \{(sI - A) V_r h_r(s) - b\} &= (s T_r - W_r^T A V_r) h_r(s) - W_r^T b \\ &= T_r \left\{ \left(sI - \widehat{H}_r\right) h_r(s) - T_r^{-1} W_r^T b \right\}. \end{aligned}$$

Part 1 of the corollary is readily established, since T_r is nonsingular. Part 2 is verified in a similar way. ∎

The approximation to $f(s)$ is then given by $c V_r (sI - \widehat{H}_r)^{-1} T_r^{-1} W_r^T b$, which is $f_{r,1}(s)$ as defined in (15). Since $f_{r,1}(s)$ and $f_{r,2}(s)$ may be computed via the application of Galerkin conditions, it is natural to question whether the terms in (15) and (17) satisfy certain "Arnoldi-like" equations. The development that follows answers this question affirmatively and shows that this process naturally fits into an implicit restart framework.

Suppose that $Q_{R_\perp}$ is the orthogonal completion of Q_R so that $[Q_R \;\; Q_{R_\perp}] \in \mathbb{R}^{m \times m}$ is an orthogonal matrix. Postmultiplying (5) by $[Q_R \;\; Q_{R_\perp}]$ enables one to express (4) and (5) as

$$\begin{aligned} b &= V_m [Q_R \;\; Q_{R_\perp}] [Q_R \;\; Q_{R_\perp}]^T l_m, \\ A V_m [Q_R \;\; Q_{R_\perp}] &= V_m [Q_R \;\; Q_{R_\perp}] [Q_R \;\; Q_{R_\perp}]^T H_m [Q_R \;\; Q_{R_\perp}] \\ &\quad + \widetilde{V}_m \widetilde{H}_m [Q_R \;\; Q_{R_\perp}], \end{aligned} \tag{18}$$

respectively, which leads to

$$b = V_r l_r + [V_m Q_{R_\perp} \;\; \widetilde{V}_m] \begin{bmatrix} Q_{R_\perp}^T l_m \\ 0 \end{bmatrix}, \tag{19}$$

$$AV_r = V_r H_r + [V_m Q_{R_\perp} \;\; \widetilde{V}_m] \begin{bmatrix} Q_{R_\perp}^T H_m Q_R \\ \widetilde{H}_m Q_R \end{bmatrix}, \tag{20}$$

where $l_r = Q_R^T l_m$ and $H_r = Q_R^T H_m Q_R$. Observe that (20) is the (1,1) block of (18). Similarly, suppose that $Q_{L\perp}$ is the orthogonal completion of Q_L so that $[Q_L \;\; Q_{L\perp}] \in \mathbb{R}^{m\times m}$ is an orthogonal matrix. Then postmultiplying (7) by $[Q_L \;\; Q_{L_\perp}]$ enables one to express (6) and (7) as

$$\begin{aligned} c^T &= W_m [Q_L \;\; Q_{L\perp}][Q_L \;\; Q_{L\perp}]^T k_m^T, \\ A^T W_m [Q_L \;\; Q_{L\perp}] &= W_m [Q_L \;\; Q_{L\perp}][Q_L \;\; Q_{L_\perp}]^T G_m^T [Q_L \;\; Q_{L\perp}] \\ &\quad + \widetilde{W}_m \widetilde{G}_m^T [Q_L \;\; Q_{L_\perp}], \end{aligned} \tag{21}$$

respectively, which leads to

$$c^T = W_r k_r^T + [W_m Q_{L\perp} \;\; \widetilde{W}_m] \begin{bmatrix} Q_{L_\perp}^T k_m^T \\ 0 \end{bmatrix}, \tag{22}$$

$$A^T W_r = W_r G_r^T + [W_m Q_{L_\perp} \;\; \widetilde{W}_m] \begin{bmatrix} Q_{L_\perp}^T G_m^T Q_L \\ \widetilde{G}_m^T Q_L \end{bmatrix}, \tag{23}$$

where $k_r^T = Q_L^T k_m^T$ and $G_r = Q_L^T G_m Q_L$. Observe that (23) is the (1,1) block of (21). The key observation is that despite the application of an additional oblique projection process, (19)–(23) have an Arnoldi-like structure, except for the second term in the right-hand side of (19) and (22). This leads one to conclude that we may restart the iterative process with a view to improving the approximation.

3.1. An implicit restart scheme. The objective of this section is to propose an implicit restart scheme based on the Arnoldi-like equations given in (19)–(23). In order to effect a restart, consider (19) and (20), where

$$[V_m Q_{R_\perp} \;\; \widetilde{V}_m] \in \mathbb{R}^{n\times(m-r+1)} \qquad \text{and} \qquad \begin{bmatrix} Q_{R_\perp}^T H_m Q_R \\ \widetilde{H}_m Q_R \end{bmatrix} \in \mathbb{R}^{(m-r+1)\times r}. \tag{24}$$

Suppose that the second term of (24) has $(m - r + 1)$ linearly independent rows; then the application of $(m - r)$ steps of restart would yield a basis for part of the controllability space spanned by the columns of $[V_r \;\; V_m Q_{R_\perp}]$. This basis is a rotation of V_m since $V_r = V_m Q_R$; therefore, $[V_r \;\; V_m Q_{R_\perp}]$ does not contribute to updating the reduced order model. Under such circumstances, the approximation errors would stagnate irrespective of the number of restarts employed. Therefore, for an effective restart scheme, one selects $2r < m$, which will be a standing assumption throughout this paper. Consider the QR decomposition

$$\begin{bmatrix} Q_{R_\perp}^T l_m & Q_{R_\perp}^T H_m Q_R \\ 0 & \widetilde{H}_m Q_R \end{bmatrix} = Q[\, \widetilde{l}_r \;\; \widetilde{H}_r], \tag{25}$$

where $[\, \widetilde{l}_r \;\; \widetilde{H}_r] \in \mathbb{R}^{(r+1)\times(r+1)}$ is upper triangular and $Q \in \mathbb{R}^{(m-r+1)\times(r+1)}$ is part of an orthogonal matrix. Then (19) and (20) may be expressed as

$$b = V_r l_r + \widetilde{V}_r \widetilde{l}_r = [V_r \;\; \widetilde{V}_r] \begin{bmatrix} l_r \\ \widetilde{l}_r \end{bmatrix}, \tag{26}$$

$$AV_r = V_r H_r + \widetilde{V}_r \widetilde{H}_r = [V_r \ \widetilde{V}_r] \begin{bmatrix} H_r \\ \widetilde{H}_r \end{bmatrix}, \tag{27}$$

where $\widetilde{V}_r := [V_m Q_{R_\perp} \ \widetilde{V}_m] Q \in \mathbb{R}^{n \times (r+1)}$. The following modified Gram–Schmidt process augments the Arnoldi-like equations (26) and (27) to yield a matrix $[V_m \ \widetilde{V}_m]$, which is part of an orthogonal matrix, and an $(r+1)$ upper Hessenberg matrix $[H_m^T \ \ \widetilde{H}_m^T]^T$ (i.e., for $1 \le j \le m$, $h_{j+r+2,j} = 0$). Observe that the first r columns of $[H_m^T \ \ \widetilde{H}_m^T]^T$ are already available from (27); furthermore, the first $2r+1$ columns of $[V_m \ \widetilde{V}_m]$ are also available prior to the restart. It is thus natural to consider a process which augments the existing data from dimension r to dimension m. The following is a modified Gram–Schmidt process that performs this task.

ALGORITHM 3.1 (an implicitly restarted modified Gram–Schmidt process).

- *For* $j = r+1, \ r+2, \ldots, m$,
- $w := Av_j$,
- *for* $i = 1, 2, \ldots, r+j$, $\begin{cases} h_{i,j} := w^T v_i, \\ w := w - v_i h_{i,j}, \end{cases}$
- $h_{j+r+1,j} := \|w\|_2$ *and* $v_{j+r+1} := w / h_{j+r+1,j}$.

Observe that $H_m = V_m^T A V_m$ is satisfied. Furthermore, the associated controllability Arnoldi equations in (4) and (5) remain in force with H_m, V_m, $\widetilde{H}_m$, $\widetilde{V}_m$, and $l_m = [l_r^T \ \widetilde{l}_r^T \ 0]^T$ defined by the implicitly restarted Gram–Schmidt process. These variables overwrite those computed in the previous restart step. To clarify the structure, we present an illustrative example in which $r = 2$ and $m = 6$. Writing (26) and (27) as

$$b = [v_1 \ v_2 \mid v_3 \ v_4 \ v_5] \left[\begin{array}{c} l_1 \\ l_2 \\ \hline l_3 \\ 0 \\ 0 \end{array}\right], \qquad A[v_1 \ v_2] = [v_1 \ v_2 \mid v_3 \ v_4 \ v_5] \left[\begin{array}{cc} \times & \times \\ \times & \times \\ \hline \times & \times \\ \times & \times \\ 0 & \times \end{array}\right],$$

the following structure is obtained by augmenting these equations to $m = 6$:

$$b = [v_1 \ v_2 \ v_3 \ v_4 \ v_5 \ v_6] \begin{bmatrix} l_1 \\ l_2 \\ l_3 \\ 0 \\ 0 \\ 0 \end{bmatrix} =: V_m l_m,$$

$$A[v_1 \ v_2 \ v_3 \ v_4 \ v_5 \ v_6] = [v_1 \ v_2 \ v_3 \ v_4 \ v_5 \ v_6 \mid v_7 \ v_8 \ v_9] \left[\begin{array}{cccccc} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \\ \hline 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & 0 & \times \end{array}\right],$$

$$\Rightarrow \qquad AV_m = [V_m \ \widetilde{V}_m] \begin{bmatrix} H_m \\ \widetilde{H}_m \end{bmatrix}.$$

In order to effect a restart with the observability Arnoldi equations, consider the QR decomposition

$$\begin{bmatrix} Q_{L_\perp}^T k_m^T & Q_{L_\perp}^T G_m^T Q_L \\ 0 & \widetilde{G}_r^T Q_L \end{bmatrix} = U[\ \widetilde{k}_r^T\ \widetilde{G}_r^T], \tag{28}$$

where $[\ \widetilde{k}_r^T\ \widetilde{G}_r^T] \in \mathbb{R}^{(r+1)\times(r+1)}$ is upper triangular and $U \in \mathbb{R}^{(m-r+1)\times(r+1)}$ is part of an orthogonal matrix. Thus, the observability Arnoldi equations may be expressed as

$$c^T = W_r k_r^T + \widetilde{W}_r \widetilde{k}_r^T \qquad \text{and} \qquad A^T W_r = W_r G_r^T + \widetilde{W}_r \widetilde{G}_r^T, \tag{29}$$

where $\widetilde{W}_r = [W_m Q_{L\perp}\ \widetilde{W}_m]U \in \mathbb{R}^{n\times(r+1)}$ is part of an orthogonal matrix. One may now employ the implicitly restarted modified Gram–Schmidt process to augment (29) and to produce $G_m, W_m, \widetilde{W}_m$, and k_m satisfying the Arnoldi equations (6) and (7). A reduced-order model is then formed following (13) or (14). Suppose that the undesirable features present in this model may be removed by two full-rank matrices T_L and T_R such that $T_L^T T_R = I_r$, then Proposition 3.1 may be used to extract the unwanted features by effecting a further state reduction. The restart process may then be repeated until convergence. The following is an outline of the implicitly restarted model-reduction algorithm.

ALGORITHM 3.2 (implicitly restarted model-reduction algorithm).

- **Start:** *Specify m and r such that $m > 2r$.*
  1. *Perform m steps of the Arnoldi process with (A, b) to find H_m, $\widetilde{H}_m$, V_m, $\widetilde{V}_m$, and l_m.*
  2. *Perform m steps of the Arnoldi process with (A^T, c^T) to find G_m, $\widetilde{G}_m$, W_m, $\widetilde{W}_m$, and k_m.*
- **Restart:** *Effect the QR decompositions*

$$T_R = [Q_R\ Q_{R_\perp}] \begin{bmatrix} R_R \\ 0 \end{bmatrix} \qquad \text{and} \qquad (T_m)^{-T} T_L = [Q_L\ Q_{L_\perp}] \begin{bmatrix} R_L \\ 0 \end{bmatrix}.$$

  1. *Effect the QR decompositions* (25) *and* (28) *which define the terms in* (26), (27), *and* (29).
  2. *Evaluate the residual errors*; *if satisfied, form the reduced-order model using* (15) *or* (17) *and stop*; *otherwise, continue.*
  3. *Effect $m - r$ implicitly restarted, modified Gram–Schmidt steps using $[l_r^T\ \widetilde{l}_r^T]^T$, $[H_r^T\ \widetilde{H}_r^T]^T$, and $[V_r\ \widetilde{V}_r]$ to yield l_m, $[H_m^T\ \widetilde{H}_m^T]^T$, and $[V_m\ \widetilde{V}_m]$.*
  4. *Effect $m - r$ implicitly restarted, modified Gram–Schmidt steps using $[k_r\ \widetilde{k}_r]$, $[G_r\ \widetilde{G}_r]$, and $[W_r\ \widetilde{W}_r]$ to yield k_m, $[G_m\ \widetilde{G}_m]$, and $[W_m\ \widetilde{W}_m]$.*

Remark. 3.3. In practice, it is advisable to compute the QR factorizations in step 1 of the restart via the Gram–Schmidt or modified Gram–Schmidt orthogonalization processes. For full details, including the case that the left-hand sides of (25) and (28) do not have full-column rank, see [4, 13].

Two elements of Algorithm 3.2 have yet to be discussed: the first addresses the selection of appropriate T_L and T_R of the restart step; the second concerns computable error formulas for step 2 of the restart process. The second is the object of the next subsection; the first is covered in section 4.

3.2. Theoretical properties. The aim of this section is to provide a theoretical analysis of the implicit restart algorithm presented in section 3.1.

To gauge the quality of the reduced-order model as each restart is completed, the following theorem provides computable expressions for the $\mathcal{L}^\infty$-norm of the residual errors.

THEOREM 3.3. *Suppose that m steps of the Arnoldi process have been completed and that the controllability and observability Arnoldi equations are given by* (26) *and* (27) *and* (29)*, respectively. The $\mathcal{L}^\infty$-norm of the residual errors associated with the approximate solutions to* (9) *are given by*

$$\|b-(sI-A)V_r h_r(s)\|_\infty = \left\| \begin{bmatrix} T_r^{-1}W_r^T\widetilde{V}_r \\ I \end{bmatrix} \Big(\widetilde{l}_r + \widetilde{H}_r h_r(s)\Big) \right\|_\infty, \tag{30}$$

$$\big\|c-g_r(s)W_r^T(sI-A)\big\|_\infty = \left\| \Big(\widetilde{k}_r + g_r(s)\widetilde{G}_r\Big) \begin{bmatrix} \widetilde{W}_r^T V_r T_r^{-1} & I \end{bmatrix} \right\|_\infty. \tag{31}$$

Proof. Substituting (26) and (27) into the left-hand side of (30) yields

$$\begin{aligned} \|b-(sI-A)V_r h_r(s)\|_\infty &= \left\| \begin{bmatrix} l_r-(sI-H_r)h_r(s) \\ \widetilde{l}_r + \widetilde{H}_r h_r(s) \end{bmatrix} \right\|_\infty \\ &= \left\| \begin{bmatrix} -T_r^{-1}W_r^T\widetilde{V}_r\widetilde{l}_r + (H_r-\widehat{H}_r)h_r(s) \\ \widetilde{l}_r + \widetilde{H}_r h_r(s) \end{bmatrix} \right\|_\infty. \end{aligned}$$

Substituting $T_r^{-1}W_r^T\times$(27) into the (1,1) block yields the desired result. The residual error expression of (31) is derived in a similar way using (29). ∎

Next, we establish properties of the reduced-order model that are reminiscent of those derived in [18]. We begin by deriving low-rank approximate solutions to (3) by imposing Galerkin-type conditions on their associated residual errors. Suppose that the low-rank Gramians have the form $P_r = V_r X_r V_r^T$ and $Q_r = W_r Y_r W_r^T$ for some symmetric matrices X_r and $Y_r \in \mathbb{R}^{r\times r}$. The residual error functions associated with a particular choice of X_r and Y_r are then defined by

$$R_r = AV_rX_rV_r^T + V_rX_rV_r^TA^T + bb^T, \qquad S_r = A^TW_rY_rW_r^T + W_rY_rW_r^TA + c^Tc.$$

The residual error functions may be factorized by using (26)–(29) to yield

$$\begin{aligned} R_r = {} & \begin{bmatrix} V_r & (I-V_rT_r^{-1}W_r^T)\widetilde{V}_r \end{bmatrix} \\ & \times \begin{bmatrix} \widehat{H}_rX_r + X_r\widehat{H}_r^T + T_r^{-1}W_r^Tbb^TW_r(T_r^T)^{-1} & X_r\widetilde{H}_r^T + T_r^{-1}W_r^Tb\widetilde{l}_r^T \\ \widetilde{H}_rX_r + \widetilde{l}_rb^TW_r(T_r^T)^{-1} & \widetilde{l}_r\widetilde{l}_r^T \end{bmatrix} \\ & \times \begin{bmatrix} V_r^T \\ \widetilde{V}_r^T(I-W_r(T_r^T)^{-1}V_r^T) \end{bmatrix}, \end{aligned} \tag{32}$$

$$\begin{aligned} S_r = {} & \begin{bmatrix} W_r & (I-W_r(T_r^T)^{-1}V_r^T)\widetilde{W}_r \end{bmatrix} \\ & \times \begin{bmatrix} \widehat{G}_r^TY_r + Y_r\widehat{G}_r + (T_r^T)^{-1}V_r^Tc^TcV_rT_r^{-1} & Y_r\widetilde{G}_r + (T_r^T)^{-1}V_r^Tc^T\widetilde{k}_r \\ \widetilde{G}_r^TY_r + \widetilde{k}_r^TcV_rT_r^{-1} & \widetilde{k}_r^T\widetilde{k}_r \end{bmatrix} \\ & \times \begin{bmatrix} W_r^T \\ \widetilde{W}_r^T(I-V_rT_r^{-1}W_r^T) \end{bmatrix}. \end{aligned} \tag{33}$$

The Arnoldi–Lyapunov solvers considered here seek symmetric X_r and Y_r such that the residual R_r and S_r satisfy orthogonality properties with respect to parts of the

controllability and observability subspaces spanned by the columns of V_r and W_r, respectively. The following theorem determines the low-rank Gramians that satisfy such orthogonality conditions.

THEOREM 3.4. *Suppose that m steps of the Arnoldi process have been completed and that the controllability and observability Arnoldi equations are given by* (26) *and* (27) *and* (29)*, respectively. Furthermore, suppose that $\lambda_i(\widehat{H}_r) + \bar{\lambda}_j(\widehat{H}_r) \neq 0 \ \ \forall i, j$ and $\lambda_i(\widehat{G}_r) + \bar{\lambda}_j(\widehat{G}_r) \neq 0 \ \ \forall i, j$. Then*

1. *$W_r^T R_r W_r = 0$ if and only if*

$$\widehat{H}_r X_r + X_r \widehat{H}_r^T + T_r^{-1} W_r^T b b^T W_r (T_r^T)^{-1} = 0. \tag{34}$$

Under these conditions,

$$\|R_r\|_F = \left\| \begin{bmatrix} H_r X_r + X_r H_r^T + l_r l_r^T & X_r \widetilde{H}_r^T + l_r \widetilde{l}_r^T \\ \widetilde{H}_r X_r + \widetilde{l}_r l_r^T & \widetilde{l}_r \widetilde{l}_r^T \end{bmatrix} \right\|_F.$$

2. *$V_r^T S_r V_r = 0$ if and only if*

$$\widehat{G}_r^T Y_r + Y_r \widehat{G}_r + (T_r^T)^{-1} V_r^T c^T c V_r T_r^{-1} = 0. \tag{35}$$

Under these conditions,

$$\|S_r\|_F = \left\| \begin{bmatrix} G_r^T Y_r + Y_r G_r + k_r^T k_r & Y_r \widetilde{G}_r + k_r^T \widetilde{k}_r \\ \widetilde{G}_r^T Y_r + \widetilde{k}_r^T k_r & \widetilde{k}_r^T \widetilde{k}_r \end{bmatrix} \right\|_F.$$

3. *X_r and $T_r^T Y_r T_r$ are, respectively, the controllability and observability Gramians of $f_{r,1}(s)$.*

4. *$T_r X_r T_r^T$ and Y_r are, respectively, the controllability and observability Gramians of $f_{r,2}(s)$.*

Proof. The proof is similar to the proofs of Theorems 2.3 and 3.2 in [18] and is omitted. □

The following result establishes that the low-rank Gramians P_r and Q_r are the exact Gramians of a pair of perturbed Lyapunov equations. It also gives the approximations $f_{r,1}(s)$ and $f_{r,2}(s)$ defined in (15) and (17), respectively, as minimal realizations of various perturbations of $f(s)$.

THEOREM 3.5. *Suppose that m steps of the Arnoldi process have been completed and that the controllability and observability Arnoldi equations are given by* (26) *and* (27) *and* (29)*, respectively. Suppose that $P_r := V_r X_r V_r^T$ and $Q_r := W_r Y_r W_r^T$ are the low-rank approximate solutions to* (3)*, where X_r and Y_r satisfy* (34) *and* (35)*, respectively.*

1. *Define the perturbations $\Delta = (I - V_r T_r^{-1} W_r^T)$, $\Delta_1 = \Delta \widetilde{V}_r \widetilde{H}_r V_r^T$, and $\Delta_2 = W_r \widetilde{G}_r \widetilde{W}_r^T \Delta$. Then*

$$(A - \Delta_1) P_r + P_r (A - \Delta_1)^T + (I - \Delta) b b^T (I - \Delta)^T = 0, \tag{36}$$

$$(A - \Delta_2)^T Q_r + Q_r (A - \Delta_2) + (I - \Delta)^T c^T c (I - \Delta) = 0. \tag{37}$$

Furthermore,

$$\|\Delta_1\|_F^2 = \left\|\widetilde{H}_r\right\|_F^2 + \left\|T_r^{-1} W_r^T \widetilde{V}_r \widetilde{H}_r\right\|_F^2,$$

$$\|\Delta_2\|_F^2 = \left\|\widetilde{G}_r\right\|_F^2 + \left\|\widetilde{G}_r \widetilde{W}_r^T V_r T_r^{-1}\right\|_F^2,$$

$$\|\Delta\|_F^2 = n - 2r + \left\|T_r^{-1}\right\|_F^2.$$

2. *Define* $\Delta_3 := \Delta_1 + \Delta_2$. *Then* $W_r^T \Delta_1 V_r = W_r^T \Delta_2 V_r = W_r^T \Delta V_r = W_r^T \Delta_3 V_r = 0$. *Furthermore,*

$$(A - \Delta_3)P_r + P_r(A - \Delta_3)^T + (I - \Delta)bb^T(I - \Delta)^T = 0,$$
$$(A - \Delta_3)^T Q_r + Q_r(A - \Delta_3) + (I - \Delta)^T c^T c(I - \Delta) = 0.$$

3. *Define* $f_{\Delta_i}(s) \stackrel{s}{=} (\ (A - \Delta_i),\ (I - \Delta)b,\ c(I - \Delta), 0\ )$ *for* $i = 1, 2, 3$. *Then*

$$f_{r,1}(s) = f_{r,2}(s) = f_{\Delta_i}(s)\ \forall s, \qquad i = 1, 2, 3.$$

Proof. Substituting (34) into (32) and rearranging yields (36), while substituting (35) into (33) and rearranging yields (37). Part 2 follows by direct calculation using part 1. The proof of part 3 is similar to the proof of Corollary 3.3 in [18] and is omitted. ∎

The effects of Δ_3 and Δ are to perturb A, b, and c in such a way that the nonminimal modes of the perturbed system are simultaneously uncontrollable and unobservable. The perturbation Δ_1 to the transition matrix of $f(s)$ yields $n - r$ uncontrollable modes while the perturbation Δ_2 gives rise to $n - r$ unobservable modes [18].

Observe that Δ_1, Δ_2, and Δ_3 are additive perturbations on the state transition matrix of $f(s)$ while Δ is a multiplicative perturbation on the input and output vectors b and c. It is interesting to observe that despite the fact that Δ_1, Δ_2, and Δ_3 have different Frobenius norms, each perturbed linear system is a different realization of the same transfer function.

Although the Arnoldi equations (4)–(7) continue to be satisfied after any number of restarts, the structure of the variables $H_m, \widetilde{H}_m, l_m, k_m, G_m$, and $\widetilde{G}_m$ differs from that required by the Arnoldi process, namely, that

$$E_m := \begin{bmatrix} l_m & H_m \\ 0 & \widetilde{H}_m \end{bmatrix}, \qquad F_m := \begin{bmatrix} k_m & 0 \\ G_m & \widetilde{G}_m \end{bmatrix}$$

are upper and lower triangular, respectively. Instead, E_m and F_m for the restart scheme are, respectively, r-upper and r-lower Hessenberg (i.e., for $1 \le j \le m$, $(E_m)_{j+r+1,j} = (F_m)_{j,j+r+1} = 0$) (see Algorithm 3.2 and Algorithm 3.1 and the subsequent discussion). This implies that the moment-matching property

$$cA^{i-1}b = c_m A_m^{i-1} b_m, \qquad 1 \le i \le 2m \tag{38}$$

[11], which essentially follows from the triangular structure of E_m and F_m, no longer applies for the implicit restart scheme and an alternative justification of the scheme is required. In the implicitly restarted model-reduction algorithm in [14, 15], the authors give an equation similar to (38), relating modified moments of the original and restarted Lanczos model. Here, we show that (38) is still satisfied, albeit for lower values of i.

THEOREM 3.6. *Suppose that* $H_m, \widetilde{H}_m, l_m, k_m, G_m$, *and* $\widetilde{G}_m$ *are output by Algorithm* 3.2; *define*

$$A_m = H_m + T_m^{-1} W_m^T \widetilde{V}_m \widetilde{H}_m = T_m^{-1} W_m^T A V_m, \tag{39}$$
$$b_m = l_m = T_m^{-1} W_m^T b, \qquad c_m = k_m T_m = c V_m$$

for nonsingular $T_m = W_m^T V_m$; *and let* k *be the largest integer satisfying*

$$k \leq \frac{m}{r+1}. \tag{40}$$

Then

$$A^{i-1}b = V_m A_m^{i-1} b_m, \qquad cA^{i-1} = c_m A_m^{i-1} T_m^{-1} W_m^T, \qquad 1 \leq i \leq k. \tag{41}$$

Hence,

$$cA^{i-1}b = c_m A_m^{i-1} b_m, \qquad 1 \leq i \leq 2k. \tag{42}$$

Proof. It follows from the r-upper triangular structure of E_m that

$$\widetilde{H}_m H_m^{i-1} l_m = 0, \qquad 1 \leq i \leq k-1. \tag{43}$$

Hence (39) implies that

$$H_m^{i-1} l_m = A_m^{i-1} l_m = A_m^{i-1} b_m, \qquad 1 \leq i \leq k. \tag{44}$$

Repeated evaluation of $b, Ab, \ldots, A^i b$ using (4), (5), (43), and (44) verifies the first part of (41). A similar procedure, using $G_m, \widetilde{G}_m$, and k_m, verifies the second part of (41). Finally, (42) follows from (41) upon noting that

$$\begin{aligned} cA^{2i-1}b &= cA^{i-1}AA^{i-1}b = c_m A_m^{i-1} T_m^{-1} W_m^T A V_m A_m^{i-1} B_m \\ &= c_m A_m^{2i-1} b_m, \quad 1 \leq i \leq k. \end{aligned}$$

Notice that when $r = 0$ (no restarts), $k = m$ and (42) reduces to (38). ∎

Remark. 3.4. One difficult issue associated with any restart scheme is the choice of m. Clearly, if $m = n-1$, then $f_{m,1}(s) = f(s)$. So the question is how small should m be to guarantee convergence of $f_{r,1}(s)$ to the balanced truncation of $f(s)$? Theorem 3.6 states that the restarted process generating $f_{m,1}(s)$ matches fewer moments than an $f_{m,1}(s)$ based on no restarts. One interpretation is that, without restarts, $f_{m,1}(s)$ tends to be a good approximation to the high-frequency component of $f(s)$, while effecting implicit restarts via the stable projection and balanced truncation of $f_{m,1}(s)$, improves the approximation at low frequencies at the expense of degraded high-frequency behavior.

Suppose that r is given and that $f_{m,1}(s)$ must match at least a given number of moments of $f(s)$ at $s = \infty$. Then (40) and (42) suggest a minimum value of m to guarantee the moment-matching condition.

4. Stable projection and balanced truncation. The objective of this section is to suggest transformations T_L and T_R which enable Algorithm 3.2 to form stable partial realizations that retain the low-frequency characteristics of $f(s)$.

Suppose that m steps of the Arnoldi process have been completed and that $f_m(s)$ is an unstable partial realization of $f(s)$. The approach proposed here then determines T_L and T_R, which effects a stable projection of $f_m(s)$. In other words, the application of T_L and T_R to $f_m(s)$ yields $f_{m+}(s)$, where $f_m(s) = f_{m+}(s) + f_{m-}(s)$ in which $f_{m+}(s)$ is stable and $f_{m-}(s)$ is antistable. For the purposes of the present discussion, suppose that $f_m(s) = c_m(sI - A_m)^{-1} b_m$, then transform A_m to a block-ordered real Schur form

$$T_1 A_m T_1^T = A_s = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$$

in which T_1 is orthogonal, $A_{11} \in \mathbb{R}^{p\times p}$ is stable, and $A_{22} \in \mathbb{R}^{(m-p)\times(m-p)}$ is anti-stable. Following this change of basis, the system of first-order differential equations is described by

$$\dot{x} = A_s x + T_1 b_m u, \qquad y = c_m T_1^T x. \tag{45}$$

The next step of the stable projection process is to eliminate the (1,2) block of A_s by solving the Sylvester equation $A_{11}X - XA_{22} + A_{12} = 0$, which has a solution due to the inertia properties of A_{11} and A_{22}; see [8]. Applying the basis change $T_2 = \left[\begin{array}{cc} I & -X \\ 0 & I \end{array}\right]$ to the linear dynamical system in (45) yields

$$T_2 A_s T_2^{-1} =: \left[\begin{array}{cc} A_{11} & 0 \\ 0 & A_{22} \end{array}\right], \quad T_2 T_1 b_m =: \left[\begin{array}{c} b_1 \\ b_2 \end{array}\right], \quad c_m T_1^T T_2^{-1} =: \left[\begin{array}{cc} c_1 & c_2 \end{array}\right].$$

With this decomposition complete, we conclude that

$$T_{L_1} := T_1^T T_2^T \left[\begin{array}{c} I \\ 0 \end{array}\right] = T_1^T \left[\begin{array}{c} I \\ -X^T \end{array}\right] \qquad \text{and} \qquad T_{R_1} := T_1^T T_2^{-1} \left[\begin{array}{c} I \\ 0 \end{array}\right] = T_1^T \left[\begin{array}{c} I \\ 0 \end{array}\right].$$

Finally, the stable part of $f_m(s)$ is

$$f_{m+}(s) := \left[\begin{array}{c|c} T_{L_1}^T A_m T_{R_1} & T_{L_1}^T b_m \\ \hline c_m T_{R_1} & 0 \end{array}\right] = c_1 (sI - A_{11})^{-1} b_1.$$

This selection of T_L and T_R may be used in Algorithm 3.2 to yield an implicitly restarted model-reduction algorithm that is reminiscent of [14, 15].

Similar to a power method, Krylov subspace methods generate partial realizations in which the spectrum of A_m is known to approximate the outer part of the spectrum of A. The presence of such eigenvalues contributes little to the low-frequency characteristics of a dynamical model and may be removed without altering the model's behavior. It is therefore natural to consider the application of a model-reduction step to $f_{m+}(s)$ whose purpose is to extract any redundant modes that might be present. To this end, one resorts to either the square root or the Schur-based algorithms expounded in [24]. Suppose that the state dimension of $f_{m+}(s)$ is p; then the following procedure determines the additional transformations, T_{L_2} and T_{R_2}, which extract the undesirable modes.

ALGORITHM 4.1 (square root algorithm).

- *Calculate the solutions P_s and Q_s to the Lyapunov equations*

$$A_{11}P_s + P_s A_{11}^T + b_1 b_1^T = 0 \qquad \textit{and} \qquad A_{11}^T Q_s + Q_s A_{11} + c_1^T c_1 = 0.$$

- *Effect the factorizations $P_s = L_r L_r^T$ and $Q_s = L_o L_o^T$, and compute the singular value decomposition $\widehat{U}\Sigma_p \widehat{V}^T = L_o^T L_r$, where $\Sigma_p = \operatorname{diag}(\sigma_1, \ldots, \sigma_p)$ and $\sigma_1 \geq \cdots \geq \sigma_p$.*
- *Suppose that the first r modes of $f_{m+}(s)$ are to be retained. Then define the transformations $T_{L_2} = L_o \widehat{U}_r \Sigma_r^{-1/2} \in \mathbb{R}^{r\times p}$ and $T_{R_2} = L_r \widehat{V}_r \Sigma_r^{-1/2} \in \mathbb{R}^{r\times p}$, where $\Sigma_r = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ and $\widehat{U}_r$ and $\widehat{V}_r$ are the first r columns of $\widehat{U}$ and $\widehat{V}$, respectively.*

We could also use the Schur-based algorithm in [24], which would yield different T_{L_2} and T_{R_2} but would result in a different realization of the same reduced-order model. In [24], Safonov and Chiang show that the square root and Schur-based

model-reduction algorithms are equivalent to Moore's balanced truncation method [20]. In the present setting, the reduced-order model is given by

$$f_r(s) = \left[\begin{array}{c|c} T_{L_2}^T A_{11} T_{R_2} & T_{L_2}^T b_1 \\ \hline c_1 T_{R_2} & 0 \end{array}\right] =: \left[\begin{array}{c|c} A_r & b_r \\ \hline c_r & 0 \end{array}\right]. \tag{46}$$

In a practical implementation, suppose that $f_m(s)$ is an unstable partial realization that contains several redundant modes. Then, it is natural to combine the stable projection and balanced truncation processes to form composite transformations that yield an $f_r(s)$ which is both stable and free of redundant modes. Such composite projectors are given by

$$T_L = T_{L_1} T_{L_2} = T_1^T \left[\begin{array}{c} I \\ -X^T \end{array}\right] L_o \hat{U}_r \Sigma_r^{-1/2} \text{ and } T_R = T_{R_1} T_{R_2} = T_1^T \left[\begin{array}{c} I \\ 0 \end{array}\right] L_r \hat{V}_r \Sigma_r^{-1/2},$$

which may be applied to $f_m(s)$ to yield $f_r(s)$ in a single step. It is interesting to observe that an implicit restart scheme may be obtained for any T_R and T_L provided that $T_L^T T_R = I_r$.

The following corollary establishes that the oblique projection methods of this paper are closely related to balanced truncation, namely, that both methods satisfy orthogonality conditions with respect to oblique projectors.

COROLLARY 4.1. *Suppose that A_r, b_r, and c_r are defined in* (46) *and define $h_{bal}(s) = (sI - A_r)^{-1} b_r$ and $g_{bal}(s) = c_r (sI - A_r)^{-1}$. Then*

$$(sI - A_{11}) T_{R_2} h_{bal}(s) - b_1 \ \perp \ T_{L_2} \quad \text{and} \quad g_{bal}(s) T_{L_2}^T (sI - A_{11}) - c_1 \ \perp \ T_{R_2}.$$

Proof. The proof is similar to the proof of Corollary 3.2. □

5. Numerical experiments. The purpose of this section is to illustrate, with the help of two examples, the behavior of the implicitly restarted model-reduction algorithm presented in sections 3 and 4. The tests reported here were performed on a Sparc-10 Sun workstation using Pro-MATLAB 4.2 which carries out operations to a unit round off of 2.22×10^{-16}.

The first problem is set up with $A \in \mathbb{R}^{n \times n}$, where $n = 100$ and the top left-hand 4×4 block of A is set to

$$\left[\begin{array}{cccc} -0.01 & 0.1 & 0 & 0 \\ -0.1 & -0.01 & 0 & 0 \\ 0 & 0 & -0.1 & 0.5 \\ 0 & 0 & -0.5 & -0.1 \end{array}\right],$$

while the remaining nonzero elements of A are uniformly distributed in $(0, -1)$ and are all located on the leading diagonal. Consequently, all the system poles are real except for four, which are $-0.01 \pm 0.1j$ and $-0.1 \pm 0.5j$. The first 10 elements of b and c are uniformly distributed in $[0, 1]$, while the 90 remaining elements are uniformly distributed in $[0, 1/25]$. We take $m = 10$ and $r = 4$. The infinity norm of $f(s)$ may be computed from $\max|f(j\omega)| \ \forall \omega \in \mathbb{R}$. Table 1 shows the evolution of the $\mathcal{L}^\infty$-norm of the error expressions of (30) and (11), denoted here by Err2 and Err3, respectively, against the number of restarts. The first column of Table 1 shows Err1, which denotes the $\mathcal{L}^\infty$-norm of the error $f_{bal}(s) - f_{r,1}(s)$, where $f_{bal}(s)$ is the rth-order balanced truncation of $f(s)$. The table indicates that Err2 and Err3 fall in magnitude as the number of restarts increases; however, as is well known, Galerkin

TABLE 1
$\mathcal{L}^\infty$-error norms associated with the implicitly restarted Arnoldi model-reduction scheme.

Restarts	Err1	Err2	Err3	Restarts	Err1	Err2	Err3
0	.3245	.7737	.3414	8	.0032	.7284	.0563
1	.1782	.7609	.3158	9	.0026	.7284	.0491
2	.0790	.7294	.2317	10	.0021	.7284	.0493
3	.0360	.7294	.1344	11	.0018	.7284	.0463
4	.0201	.7291	.1013	12	.0015	.7285	.0474
5	.0117	.7287	.0656	13	.0012	.7285	.0457
6	.0069	.7288	.0712	14	.0009	.7285	.0469
7	.0044	.7285	.0607	15	.0007	.7285	.0456

conditions of the type in (10) do not guarantee a nonincreasing evolution of Err2 or Err3. Note that Err2 stagnates after an initial drop indicating convergence. Our experience with similar examples indicates that Err1 always tends to zero (for large enough m), which implies that $f_{r,1}(s)$ converges to $f_{bal}(s)$, although proving this remains an open question. Observe that Err2 and Err3 do not converge to zero. This follows from the fact that when approximating $f(s)$ by a stable kth-order model $f_k(s)$, $\|f(s) - f_k(s)\|_\infty$ is greater than or equal to the $(k+1)$st Hankel singular value of $f(s)$ [8].

It is known that for many restart algorithms there exists a value of m below which convergence of the solution is very slow or not possible [25]. The second example illustrates that this is also the case for the restart algorithm presented here. The problem is set up with $A \in \mathbb{R}^{n\times n}$, where $n = 300$. The eigenvalues of A are all in the open left half plane; the real parts are uniformly distributed in the interval $[-1, 0)$, while the imaginary parts are randomly distributed in the interval $[-5, 5]$. $b,\ c^T \in \mathbb{R}^{n\times 1}$ are random.

TABLE 2
Evolution of relative $\mathcal{L}^\infty$-error norms for $m = 60$.

Restarts	1	2	3	4	5	6	7	8	9
E (%)	64	79	64	68	66	123	67	87	82

TABLE 3
Evolution of relative $\mathcal{L}^\infty$-error norms for $m = 70$.

Restarts	1	2	3
E (%)	333	23	4

Tables 2 to 4 illustrate the evolution of the percentage error $E = \|f_{r,1}(s) - f_{bal}(s)\|_\infty / \|f_{bal}(s)\|_\infty$ as a function of m and the number of restarts. Here $r = 5$ and $f_{bal}(s)$ is the rth-order balanced truncation of $f(s) = c(sI - A)^{-1}b$. Note that for $m = 60$, no convergence occurs; for $m = 70$, $f_{r,1}(s)$ converges to within 4% of $f_{bal}(s)$ after three restarts; while for $m = 75$, convergence is within 0.1% after only two restarts. This example is typical of our numerical experience. Determining a least value of m to guarantee convergence is still an unresolved problem and is under investigation (see Remark 3.4).

6. Conclusions. This paper presents and tests a model-reduction algorithm for large-scale, stable, linear, and time-invariant dynamic systems. We have developed a

TABLE 4
Evolution of relative $\mathcal{L}^\infty$-error norms for $m = 75$.

Restarts	1	2
E (%)	4.7	0.1

technique which combines the oblique Krylov subspace projectors with further projectors in order to obtain stable reduced-order approximate models that are able to approximate the low-frequency behavior of the dynamic system. We also established that this technique fits naturally within an implicit restart framework that defines an iterative procedure able to refine approximations. Exact low-dimensional expressions for the $\mathcal{L}^\infty$-norm of the residual errors are also derived. Our numerical experiments on several large-scale examples indicate that this process converges to the balanced truncation of the dynamic system; however, formally establishing this claim remains an open research problem. In place of the Arnoldi process, one may equally employ the Lanczos algorithm to derive an implicitly restarted scheme for stable partial realization. The derivation of such a scheme follows similar lines to those presented in this paper, except that in the Lanczos setting the biorthogonality of $[V_m \quad \widetilde{V}_m]$ and $[W_m \quad \widetilde{W}_m]$ is enforced [5, 21].

REFERENCES

[1] M. M. M. AL-HUSARI, B. HENDEL, I. M. JAIMOUKHA, E. M. KASENALLY, D. J. N. LIMEBEER, AND A. PORTONE, *Vertical stabilisation of Tokamak plasmas*, 30th Conference on Decision and Control, England, 1991.

[2] W. ARNOLDI, *The principle of minimised iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[3] D. L. BOLEY, *Krylov space methods on state-space control models*, CS Technical report, TR92-18, Department of Computer Science, University of Minnesota, Minnesota, 1992.

[4] D. L. BOLEY AND G. H. GOLUB, *The Lanczos-Arnoldi algorithm and controllability*, Systems Control Lett., 4 (1984), pp. 317–327.

[5] D. L. BOLEY AND G. H. GOLUB, *The nonsymmetric Lanczos algorithm and controllability*, Systems Control Lett., 16 (1991), pp. 97–105.

[6] P. FELDMANN AND R. W. FREUND, *Efficient Linear Circuit Analysis by Padé Approximation via the Lanczos Process*, Technical report No. 11274-940217-02TM, AT&T Bell Laboratories, Murray Hill, NJ, 1994.

[7] C. W. GEAR AND Y. SAAD, *Iterative solution of linear equations in ODE codes*, SIAM. J. Sci. Statist. Comput., 4 (1983), pp. 583–601.

[8] K. GLOVER, *All optimal Hankel norm approximations of linear multivariable systems and their $\mathcal{L}^\infty$ error bounds*, Inter. J. Control, 39 (1984), pp. 1115–1193.

[9] K. GLOVER, *Multiplicative approximation of linear multivariable systems with $\mathcal{L}_\infty$ error bounds*, American Control Conference, Seattle, WA, pp. 1705–1709, 1986.

[10] M. GREEN AND D. J. N. LIMEBEER, *Linear Robust Control*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[11] W. B. GRAGG AND A. LINDQUIST, *On the partial realization problem*, Linear Algebra Appl., 50 (1983), pp. 277–319.

[12] G. H. GOLUB, B. KÅGSTRÖM, AND P. VAN DOOREN, *Direct block tridiagonalisation of single-input single-output systems*, System Control Lett., 18 (1992), pp. 109–120.

[13] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., John Hopkins University Press, Baltimore, MD, 1990.

[14] E. J. GRIMME, D. C. SORENSEN, AND P. VAN DOOREN, *Stable partial realizations via an implicitly restarted Lanczos method*, American Control Conference, Baltimore, MD, 1994.

[15] K. GALLIVAN, E. J. GRIMME, AND P. VAN DOOREN, *Asymptotic waveform evaluation via a restarted Lanczos method*, Appl. Math. Lett., 7 (1994), pp. 75–80.

[16] I. M. JAIMOUKHA, E. M. KASENALLY, AND D. J. N. LIMEBEER, *Numerical solution of large scale Lyapunov equations using Krylov subspace methods*, 31st Conference on Decision

and Control, Tucson, AZ, 1992.
[17] I. M. Jaimoukha and E. M. Kasenally, *Krylov subspace methods for solving large Lyapunov equations*, SIAM J. Numer. Anal., 31 (1994), pp. 227–251.
[18] I. M. Jaimoukha and E. M. Kasenally, *Oblique projection methods for large scale model reduction*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 602–627.
[19] E. M. Kasenally, *Analysis of some Krylov subspace methods for large matrix equations*, IRC-PSE report, No C93-14, Imperial College, Univeristy of London, 1992.
[20] B. C. Moore, *Principal component analysis in linear systems: controllability, observability and model reduction*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 17–31.
[21] B. Parlett, *Reduction to tridiagonal form and minimal realizations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 567–593.
[22] Y. Saad, *Analysis of some krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 29 (1991), pp. 209–228.
[23] Y. Saad, *Numerical solution of large Lyapunov equations*, in Signal Processing, Scattering, Operator Theory and Numerical Methods, M. A. Kaashoek, J. H. Van Schuppen, and A. C. M. Ran, eds., Birkhäuser, Boston, Cambridge, MA, pp. 503–511, 1990.
[24] M. G. Safonov and R. Y. Chiang, *A Schur method for balanced-truncation model reduction*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 729–733.
[25] D.C. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385.
[26] P. Van Dooren, *Numerical linear algebra techniques for large scale matrix problems in systems and control*, 31st Conference on Decision and Control, Tucson, AZ, 1992.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 3, pp. 653–692, July 1997

008

A GEOMETRIC APPROACH TO PERTURBATION THEORY OF MATRICES AND MATRIX PENCILS. PART I: VERSAL DEFORMATIONS*

ALAN EDELMAN†, ERIK ELMROTH‡, AND BO KÅGSTRÖM‡

Abstract. We derive versal deformations of the Kronecker canonical form by deriving the tangent space and orthogonal bases for the normal space to the orbits of strictly equivalent matrix pencils. These deformations reveal the local perturbation theory of matrix pencils related to the Kronecker canonical form. We also obtain a new singular value bound for the distance to the orbits of less generic pencils. The concepts, results, and their derivations are mainly expressed in the language of numerical linear algebra. We conclude with experiments and applications.

Key words. Jordan canonical form, Kronecker canonical form, generalized Schur decomposition, staircase algorithm, versal deformations, tangent and normal spaces, singularity theory, perturbation theory

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

PII. S0895479895284634

Notation.

$\|x\|$	The 2-norm of a vector x.
A	A square matrix of size $n \times n$. I or I_n is the identity matrix.
A^T	The transpose of A.
A^H	The conjugate transpose of A.
$\overline{A}$	The conjugate of A.
$\|A\|_E$	The Frobenius (or Euclidean) matrix norm.
$\sigma_{\min}(A)$	The smallest singular value of A.
$\mathrm{vec}(A)$	An ordered stack of the columns of a matrix A from left to right.
$\det(A)$	Determinant of A.
$\mathrm{tr}(A)$	Trace of A.
$\ker(A)$	Kernel of space spanned by the columns of A.
$\mathrm{range}(A)$	Range of space spanned by the columns of A.
$\mathrm{diag}(A_1, \ldots, A_b)$	A block diagonal matrix with diagonal blocks A_i.
$A \otimes B$	The Kronecker product of two matrices A and B whose (i,j)th block element is $a_{ij}B$.
$A - \lambda B$	A matrix pencil of size $m \times n$.
λ_i	Eigenvalue of A or $A - \lambda B$. Also, γ_i and α are used to denote an eigenvalue.

*Received by the editors April 17, 1995; accepted for publication (in revised form) by P. Van Dooren July 12, 1996.

http://www.siam.org/journals/simax/18-3/28463.html

†Department of Mathematics, Room 2-380, Massachusetts Institute of Technology, Cambridge, MA 02139 (edelman@math.mit.edu). The research of this author was supported by NSF grant DMS-9120852 and an Alfred P. Sloan Foundation Research Fellowship.

‡Department of Computing Science, Umeå University, S-901 87 Umeå, Sweden (elmroth@cs.umu.se, bokg@cs.umu.se). The research of these authors was supported in part by Swedish National Board of Industrial and Technical Development grant 89-02578P and by Swedish Research Council for Engineering Sciences grant TFR 222-95-34.

r, r_i, s_i	$A - \lambda B$ has r distinct eigenvalues λ_i of algebraic multiplicity r_i. The sizes of the Jordan blocks associated with an eigenvalue are $s_1 \geq s_2 \geq \cdots \geq s_{r_i}$.
$J_j(\gamma_i)$	Jordan block of size $j \times j$ associated with γ_i.
$J_j(\gamma_i, \bar{\gamma}_i)$	Real Jordan block of size $2j \times 2j$ associated with a complex conjugate pair of eigenvalues.
N_j	Jordan block of size $j \times j$ associated with the infinite eigenvalue.
L_j	Singular block of right (column) minimal index of size $j \times (j+1)$.
L_j^T	Singular block of left (row) minimal index of size $(j+1) \times j$.
$\langle A - \lambda B, C - \lambda D \rangle$	Frobenius inner product of two matrix pairs.
orbit(A)	The set of matrices similar to A.
orbit$(A - \lambda B)$	The set of matrix pencils equivalent to $A - \lambda B$.
tan$(A - \lambda B)$	Tangent space of orbit$(A - \lambda B)$ at $A - \lambda B$.
nor$(A - \lambda B)$	Normal space of orbit$(A - \lambda B)$ at $A - \lambda B$.
$\mathcal{S} \oplus \mathcal{T}$	Direct sum of subspaces $\mathcal{S}$ and $\mathcal{T}$ of $\mathbf{R}^n$.
$\mathcal{S}^\perp$	Subspace perpendicular to $\mathcal{S}$. $\mathcal{S} \oplus \mathcal{S}^\perp$ is the complete space.
dim$(\mathcal{S})$	Dimension of subspace $\mathcal{S}$. dim(S) denotes dimension of subspace spanned by the columns of S.
cod$(\mathcal{S})$	Codimension is the dimension of the subspace complementary to $\mathcal{S}$.
$\mathcal{P}$	The $2mn$-dimensional space of $m \times n$ matrix pencils, i.e., $\mathcal{P} =$ tan$(A - \lambda B) \oplus$ nor$(A - \lambda B)$.
$\mathcal{V}(p)$	Deformation or (mini)versal deformation with parameter vector $p \in \mathbf{R}^l$, where $l \geq 1$. $\mathcal{V}(p)$ is also written $\mathcal{V}(p_1, p_2, \ldots, p_l)$. q is also used as parameter vector.
$Z(p)$	Deformation that spans the orthogonal complement of the orbit of a matrix A.
$Z_A(p) - \lambda Z_B(p)$	Deformation that spans the orthogonal complement of the orbit of a pencil $A - \lambda B$. Often abbreviated $Z_A - \lambda Z_B$.

1. Introduction and examples.

1.1. Introduction. Traditionally, canonical structure computations take as their input some mathematical object, a matrix or a pencil, say, and return an equivalent object that is perhaps simpler or makes clear the structure of the equivalence relation. Some example equivalence relations and corresponding canonical forms are as follows.

Structure	**Equivalence relation**	**Canonical form**
Square matrices	$A \sim X^{-1}AX$	Jordan canonical form
Rectangular matrices	$A \sim UAV$	Singular values
Rectangular matrices	$A \sim XA$	Reduced echelon form
Matrix pencils	$A - \lambda B \sim P^{-1}(A - \lambda B)Q$	Kronecker canonical form
Analytic real functions	$f(x) \sim f(\phi(x))$	$\pm x^k$

In the first three examples the input is a matrix. In the next example, the input is a pencil. In these cases, X, P, and Q are presumed nonsingular and U and V are presumed orthogonal. We presume the real functions f are analytic in a neighborhood

of zero, $f(0) = 0$, $\phi(0) = 0$, and $\phi(x)$ is monotonic and analytic near zero.

Canonical forms appear in every branch of mathematics. A few examples from control theory may be found in [21, 20, 27, 19]. However, researchers in singularity theory have asked what happens if you have not one object that you want to put into a normal form, but rather a whole family of objects nearby some particular object and you wish to put each member of the family into a canonical form in such a way that the canonical form depends smoothly on the deformation parameters.

For example, one may have a one-parameter matrix deformation of A which is simply an analytic function $\mathcal{V}(p)$ for which $\mathcal{V}(0) = A$. An n parameter deformation is defined the same way, except that $p \in \mathbf{R}^n$. Similarly, one may have n parameter deformations of pencils or functions. Remaining with the matrix example, we say two deformations $\mathcal{V}_1(p)$ and $\mathcal{V}_2(p)$ are equivalent if $\mathcal{V}_1(p)$ and $\mathcal{V}_2(p)$ have the same Jordan canonical form for each and every p. A deformation of a matrix is said to be versal if, loosely speaking, it captures all possible Jordan form behaviors near the matrix. A deformation is said to be miniversal if it does so with as few parameters as possible. A more formal discussion of these definitions may be found in section 2.

The derivation of versal and miniversal deformations requires a detailed understanding of the perturbation theory of the objects under study. In particular, one needs to understand the tangent space of the equivalence relation and how it is embedded in the entire space. In section 2, we explain the mechanics of this perturbation theory.

While we believe that versal deformations are interesting mathematical objects, this work differs from others on the subject in that our primary goal is not so much the versal deformation or the miniversal deformation, but rather the perturbation theory and how it influences the computation of the Kronecker canonical form. As such, we tend to be interested more in metrical information than topological information. Therefore, we obtain new distance formulas to the space of less generic matrix pencils in section 4. In section 5, we derive an explicit orthogonal basis for the normal space of a Kronecker canonical form. For us a versal decomposition will be an explicit decomposition of a perturbation into its tangential and normal components, and we will not derive any miniversal deformations that may have simpler forms, but hide the metric information.

Versal deformations for function spaces are discussed in [18, 25, 4, 5]. The first application of these ideas for the matrix Jordan canonical form is due to Arnold [1]. Further references closely related to Arnold's matrix approach are [30] and [6]. The latter reference also includes applications to differential equations. Applications of the matrix idea toward an understanding of companion matrix eigenvalue calculations may be found in [13]. The only other work that we are aware of that considers versal deformations of the Kronecker canonical form is by Berg and Kwatny [3], who independently derived some of the normal forms considered in this paper.

Our section 2 contains a thorough explanation of versal deformations from a linear algebra perspective. Section 3 briefly reviews matrix pencils and canonical forms. Section 4 derives the geometry of the tangent and normal spaces to the orbits of matrix pencils. Section 5 derives the versal deformations, while section 6 gives applications and illustrations.

Notation is introduced and defined the first time it appears in the text. Some (but not all) of the notation used in the paper is summarized on the previous page. For example, the glossary of Toeplitz and Hankel matrices (section 5.2) is not repeated there. Moreover, the definitions of different canonical forms (companion, Jordan,

Kronecker, generalized Schur, etc.) are introduced in their context.

1.2. Geometry of matrix space. Our guiding message is very simple: matrices should be seen in the mind's eye geometrically as points in n^2-dimensional space. A perfect vision of numerical computation would allow us to picture computations as moving matrices from point to point or manifold to manifold.

Abstractly, it hardly matters whether a vector is a column of numbers or a geometric point in space. However, without the interplay of these two representations, numerical linear algebra would not be the same. Imagine explaining without the geometric viewpoint how Householder reflections transform vectors.

In contrast, in numerical linear algebra we all know that matrices are geometric points in n^2-dimensional space, but it is rare that we actually *think* about them this way. Most often, matrices are thought of as either (sparse or dense) arrays of numbers, or they are operators on vectors.

The Eckart–Young (or Schmidt–Mirsky theorem) [29, p. 210] gives a feel for the geometric approach. The theorem states that the smallest singular value of A is the Frobenius distance of A to the set of singular matrices. One can not help but see a blob representing the set of singular matrices. This amorphous blob is most often thought of as an undesirable part of town, so unfortunately numerical analysts hardly ever study the set itself. Algebraic geometers recognize the singular matrices as a variety, meaning that the set can be defined as the zero set of a polynomial system (namely, $\det(A) = 0$). It can also be "stratified" as the union of manifolds. The most generic singular matrices are the ones with rank $n - 1$. These matrices form a manifold.

Demmel helped pioneer the development of geometric techniques [7] for the analysis of ill conditioning of numerical analysis problems. Shub and Smale [28] are applying geometrical approaches toward the solution of polynomial systems.

We believe that if only we could better understand the geometry of matrix space, our knowledge of numerical algorithms and their failures would also improve. A general program for numerical linear algebra, then, is to transfer from pure mathematicians the technology to geometrically understand the high dimensional objects that arise in numerical linear algebra. This program may not be easy to follow. A major difficulty is that pure mathematicians pay a price for their beautiful abstractions—they do not always possess a deep understanding of the individual objects that we wish to study. This makes technology transfer difficult. Even when the understanding exists somewhere, it may be difficult to recognize or may be buried under a heavy layer of notation. This makes technology transfer time consuming. Finally, even after expending time excavating, the knowledge may still be difficult to apply toward the understanding or the improving of practical algorithms. This makes technology transfer from pure mathematics frustrating.

Nevertheless, our goal as researchers is the quest for understanding which we may then apply. In this paper, we follow our program for the understanding of the Jordan and Kronecker canonical forms of matrices and matrix pencils, respectively. Many of the ideas in this paper have been borrowed from the pure mathematics literature with the goal of simplifying and applying them to the needs of numerical linear algebraists.

While this is quite a general program for numerical linear algebra, this paper focuses on a particular goal. We analyze *versal deformations* from the numerical linear algebra viewpoint and then compute normal deformations for the Kronecker canonical form. We consider both of these as stepping stones toward the far more difficult goal of truly understanding and improving staircase algorithms for the Jordan

or Kronecker canonical form. These are algorithms used in systems and control theory. The structures of these matrices or pencils reflect important physical properties of the systems they model, such as controllability [10, 32].

The user chooses a parameter η to measure any uncertainty in the data. The existence of a matrix or pencil with a different structure within distance η of the input means that the actual system may have a different structure than the approximation supplied as input. These algorithms try to perturb their input by at most η so as to find a matrix or pencil with as high a codimension as possible. The algorithm is said to *fail* if there is another perturbation of size at most η which would raise the codimension even further. Therefore, we must understand the geometry of matrix space to begin to understand how we can supply the correct information to the user. With this information, we believe that we would then be able to not only correctly provide the least generic solutions, but also understand how singularities hinder this process. Bad solutions may then be refined so as to obtain better solutions. As the next section illustrates, the geometry directly affects the perturbation theory.

1.3. Motivation: A singular value puzzle. Consider the following four nearly singular matrices:

$$M_1 = \begin{pmatrix} 0 & 1+\epsilon \\ 0 & 0 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 0 & 1 \\ \epsilon & 0 \end{pmatrix}, \quad M_3 = \begin{pmatrix} \epsilon & 1 \\ 0 & -\epsilon \end{pmatrix}, \quad M_4 = \begin{pmatrix} \epsilon & 1 \\ 0 & \epsilon \end{pmatrix}. \tag{1.1}$$

Each of these matrices are distance $O(\epsilon)$ from the Jordan block

$$J_2(0) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

What is the smaller of the two singular values of each of M_1, M_2, M_3, and M_4? The answer is

$$\sigma_{\min}(M_1) = 0, \quad \sigma_{\min}(M_2) = \epsilon, \quad \sigma_{\min}(M_3) \approx \epsilon^2, \quad \text{and} \quad \sigma_{\min}(M_4) \approx \epsilon^2.$$

A quick way to verify this algebraically is to notice that the larger singular value of each matrix is approximately 1 so that the smaller is approximately the (absolute) determinant of the matrix. Another approach that bounds the smallest singular value is the combination of the Eckart–Young theorem and the observation that these matrices are singular:

$$M_1' = M_1, \qquad M_2' = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad M_3' = \begin{pmatrix} \epsilon & 1 \\ -\epsilon^2 & -\epsilon \end{pmatrix}, \quad M_4' = \begin{pmatrix} \epsilon & 1 \\ \epsilon^2 & \epsilon \end{pmatrix}.$$

When $\epsilon = 0$ in (1.1) our four matrices become the singular 2×2 Jordan block $J_2(0)$. As ϵ varies from 0 each of the four forms in (1.1) traces out a line in matrix space. The geometric issue that is interesting here is that the line of matrices traced out as ϵ varies is {1:In, 2:Normal, 3:Tangent, 4:Tangent} to the set of singular matrices. Somehow, this feels like the "right" explanation for why the smaller singular values are $\{1{:}0,\ 2{:}\epsilon,\ 3{:}\approx \epsilon^2,\ 4{:}\approx \epsilon^2\}$.

Let us take a closer look at the set of singular matrices. The four parameters found in a 2×2 matrix M are best viewed in a transformed coordinate system:

$$M = (x, y, z, w) = x \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + y \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} + z \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + w \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} w+z & x \\ y & w-z \end{pmatrix}.$$

In this coordinate system, the singular matrices fall on the surface described by the equation $w^2 = z^2 + xy$. This is a three-dimensional surface in four-dimensional space. The traceless singular matrices ($w = 0$) fall on the cone $z^2 + xy = 0$ in three-dimensional space.

Our matrix $J_2(0)$ may now be represented as $(1,0,0,0)$ and the four lines of matrices mentioned above are

$$l_1 = \{(1+\epsilon,0,0,0)\} = \left\{ \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + \epsilon \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \right\},$$

$$l_2 = \{\quad (1,\epsilon,0,0) \quad\} = \left\{ \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + \epsilon \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \right\},$$

$$l_3 = \{\quad (1,0,\epsilon,0) \quad\} = \left\{ \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + \epsilon \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \right\},$$

$$l_4 = \{\quad (1,0,0,\epsilon) \quad\} = \left\{ \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + \epsilon \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right\}.$$

The lines l_1, l_2, and l_3 are all traceless; i.e., the matrices on each of these lines may be viewed in the three-dimensional space of the cone. The line l_1 is not only tangent to the cone, but in fact it lies in the cone. The line l_3 is tangent to one of the circular cross sections of the cone.

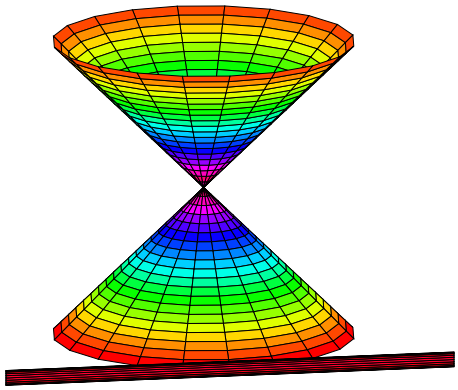


FIG. 1.1. *Cone of traceless singular matrices with "stick" representing a tangent.*

Figure 1.1 illustrates l_3 as a "stick" resting near the bottom of the cone. The line l_1 is a thin line on the cone through the same point.

The line l_4 is normal to the cone, but it is also tangent to the variety of singular matrices. One way to picture this in three dimensions is to take the three-dimensional slice of $\{w^2 = z^2 + xy\}$ corresponding to $x = 1$, i.e., $\{w^2 - z^2 = y\}$. This is a hyperboloid with the Jordan block as a saddle point. The line is the tangent to the parabola $w^2 = y$ which rests in the plane $z = 0$. Figure 1.2 illustrates this line with a cylindrical stick whose central axis is the tangent. Finally, the line l_2 is normal to the set of singular matrices.

If we move a distance ϵ away from a point on a surface along a tangent, our distance to the surface remains $O(\epsilon^2)$. This is what the singular value corresponding

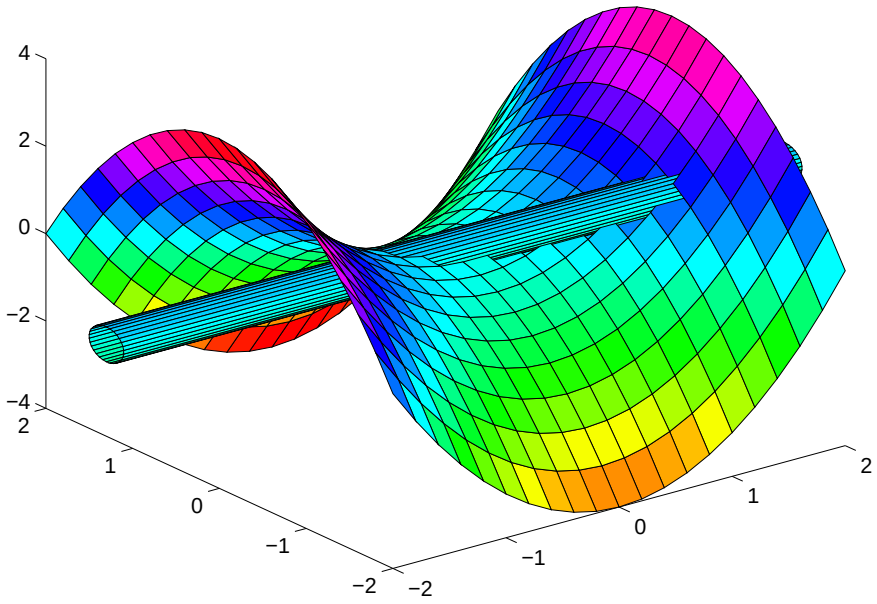


FIG. 1.2. *Variety of singular matrices. The axis of the cylindrical stick is tangent to the singular variety.*

to l_3 and l_4 is telling us. Alternatively, if we move normal to the surface as in l_2, the singular value changes more rapidly: $O(\epsilon)$.

The cone of singular matrices with $w = 0$ is not only a slice of a large dimensional space, but it is also the (closure of) the set of matrices similar to $J_2(0)$ (which we denote orbit$(J_2(0))$ in section 2.4). The matrices similar to $J_2(0)$ are singular and traceless. In fact, the only matrix that is singular and traceless that is not similar to $J_2(0)$ is the 0 matrix which is the vertex of the cone. We further explore this case in section 2.5 after we have defined versal deformations.

We conclude that the geometry of the orbit and, in particular, the directions of the tangents and normals to the orbit directly influence the eigenvalue perturbation theory.

2. Introduction to versal deformations. This introduction is designed to be readable for general audiences, but we particularly target the numerical linear algebra community.

The ideas here may be thought of as a numerical analyst's viewpoint on ideas that were inspired by Arnold's work [1] on versal deformations of matrices. Further elaboration upon Arnold's versal deformations of matrices may be found in [6, Chapters 2.9 and 2.10] and [30]. These ideas fit into a larger context of differential topology and singularity theory. Bruce and Giblin [5] have written a wonderfully readable introduction to singularity theory emphasizing the elementary geometrical viewpoint. After reading this introduction, it is easy to be lulled into the belief that one has mastered the subject, but a more advanced wealth of information may be found in [18, 25, 4]. Finally, what none of these references do very well is clearly explain that there is still much in this area that mankind does not yet fully understand.

Singularity theory may be viewed as a branch of the study of curves and surfaces, but its crowning application is toward the topological understanding of functions and their behavior under perturbations. Of course, numerical analysts are very interested in perturbations as well.

2.1. Characteristic polynomials give the "feel" of versal deformations. Let $\mathcal{V}(p)$ be a differentiable one-parameter family of matrices through $A \equiv \mathcal{V}(0)$. This is just a curve in matrix space. If A has a complicated Jordan canonical form, then very likely the Jordan canonical form of $\mathcal{V}(p)$ is a discontinuous function of p. (The Jordan canonical form, you will remember, can have nasty ones popping up unexpectedly on the superdiagonal.) It is even more desirable if that function can somehow describe the kinds of matrices that are near A.

Discontinuities are as unpleasant for pure mathematicians as they are for computers. Therefore, Arnold [1] asks what kinds of functions of p are differentiable (or many times differentiable, or analytic).

One function that comes to mind is the characteristic polynomial $\det(\mathcal{V}(p) - \lambda I)$. The coefficients of $\det(\mathcal{V}(p) - \lambda I)$ are clearly differentiable functions of p no matter how complicated a Jordan canonical form the matrix A might have. In numerical linear algebra, we never compute the characteristic polynomial because the eigenvalues are often very poorly determined by the coefficients of the characteristic polynomial. Mathematically, the characteristic polynomial is a nice function of a matrix because its coefficients, unlike the eigenvalues of the matrix, are analytic functions of the entries of the matrix.

The characteristic polynomial is a reasonable representation for the Jordan canonical form under the special circumstance that every matrix $\mathcal{V}(p)$ is *nonderogatory* (i.e., each matrix has exactly one Jordan block for each distinct eigenvalue). By a reasonable representation, we mean here that it actually encodes the Jordan canonical form of A. Theoretically, if you know the characteristic polynomial, then you know the eigenvalues with appropriate multiplicities. It follows that there is a unique nonderogatory Jordan canonical form (see Wilkinson [35, pp. 11–16 or Note 55, p. 408]). To repeat, there is a one-to-one correspondence among the n eigenvalues of a nonderogatory matrix, the characteristic polynomial of a nonderogatory matrix, and the Jordan canonical form of a nonderogatory matrix, but only the characteristic polynomial is a differentiable function of the perturbation parameter p. (The eigenvalues themselves can have first-order perturbations with the nondifferentiable form $p^{1/n}$, for example, for an $n \times n$ matrix A with only one Jordan block $J_n(\lambda)$. This is a well-known example.)

In the language of numerical linear algebra, we would say that a nonderogatory matrix A may be written in companion matrix form KCK^{-1} in such a way that differentiable perturbations to the matrix A lead to differentiable perturbations to the companion matrix C. Here the matrix K is a Krylov matrix (see [17, p. 369]). Equivalently, first-order perturbations to the matrix A are manifested as first-order perturbations to the companion matrix C. When A is a companion matrix, this gives a first-order perturbation theory for the characteristic polynomials of nearby matrices. This perturbation theory is computed in [13].

Our story would almost stop here if we were interested only in the Jordan form of nonderogatory matrices. We say "almost" because it would be a shame to stop here without explaining the ideas geometrically. Even if we did not discuss the geometry, we have reasons to continue on, since matrix space is enriched with the derogatory matrices, and also we wish to generalize these ideas about the Jordan canonical form

to cover the more complicated case of the Kronecker canonical form.

2.2. The rational canonical form is not enough for derogatory matrices. In the previous section we saw that n parameters were sufficient to specify the Jordan canonical form of any matrix in a small neighborhood of a nonderogatory matrix. What happens if the matrix is derogatory? One obvious guess turns out to be wrong. The usual generalization of the companion matrix form for derogatory matrices is the rational canonical form. If A is derogatory, it may be put in rational canonical form. This form may be thought of as the direct sum of companion matrices C_i with dimension $m_1 \geq m_2 \geq \cdots \geq m_k$. The characteristic polynomial of each C_i divides the characteristic polynomial of all the preceding $C_j, j < i$. Can any nearby matrix be expressed as the direct sum of companion matrices with dimension $m_1, m_2, \ldots, m_k$ in a nice differentiable manner? The answer is generally no; though good enough to specify the Jordan canonical form of a matrix, the rational canonical form fails to be powerful enough to specify the Jordan canonical forms of all matrices in a neighborhood. This is because there are just not enough parameters in the rational canonical form to cover all the possibilities. To have enough parameters we need a "versal deformation."

One simple example is the identity matrix (or the zero matrix). The rational canonical form has $m_1 = \cdots = m_n = 1$. The matrices with this form are the diagonal matrices, and hence every one of them is nondefective (diagonalizable). However, with an arbitrarily small perturbation of the identity, it is possible to obtain defective matrices. The rational canonical form has n parameters, which are not enough.

2.3. Versal deformation: The linearized theory. The "linearized" picture of a versal deformation is easy to understand. We therefore explain this picture before plunging into the global point of view. The general case may be nonlinear, but the linearized theory is all that really matters. For simplicity we assume that we are in real n-dimensional Euclidean space, but this assumption is not so important.

We recall the elementary fact that if $\mathcal{S}$ and $\mathcal{T}$ are subspaces of $\mathbf{R}^n$ such that $\mathcal{S} \oplus \mathcal{T} = \mathbf{R}^n$, then there exist linear projections $\pi_{\mathcal{S}}$ and $\pi_{\mathcal{T}}$ that map onto $\mathcal{S}$ and $\mathcal{T}$, respectively.

Consider a point $x \in \mathcal{S}$. We will investigate all possible perturbations y of x, but we will not be concerned with perturbations that are within $\mathcal{S}$ itself. Psychologically, we consider all the vectors in $\mathcal{S}$ to somehow be the same, so there will be no need to distinguish them. Let $\mathcal{T}$ be any linear subspace such that $\mathcal{S} \oplus \mathcal{T} = \mathbf{R}^n$; i.e., any vector may be written as the sum of an element of $\mathcal{T}$ and an element of $\mathcal{S}$ (not necessarily uniquely). Clearly if $t_1, \ldots, t_k$ span $\mathcal{T}$, then our perturbed vector $x+y$ may be written as

$$x + y = x + \sum_{i=1}^{k} p_i t_i + (\text{something in } \mathcal{S}),$$

where the p_i may be chosen as linear functions of y. We see here what will turn out to be the key idea of a versal deformation—every perturbation vector may be expressed in terms of the p_i and vectors that we are considering to all be equivalent.

We now formally introduce the local picture of versal deformations.

DEFINITION 2.1. *A* linear deformation *of the point x is a function defined on $p \in \mathbf{R}^l$:*

$$\mathcal{V}(p) = x + Tp,$$

where $T = [t_1, t_2, \ldots, t_l]$ are arbitrary directions.

The choice of the word "deformation" is meant to convey the idea that we are looking at small values of the p_i, and these perturbations are small deformations of the starting point x.

DEFINITION 2.2. *A linear deformation $\mathcal{V}_1(p)$ of the point x is* versal *if for all linear deformations $\mathcal{V}_2(q)$ of the point x, it is possible to write*

$$\mathcal{V}_2(q) = \mathcal{V}_1(\phi(q)) + \theta(q),$$

where $\phi(q)$ is a linear function from $q_1, \ldots, q_m$ to $p_1, \ldots, p_l$ with $\phi(0) = 0$ and θ is a linear function from q into $\mathcal{S}$ with $\theta(0) = 0$.

We now explain why $\mathcal{V}_1(p) = x + \sum_{i=1}^{l} p_i t_i$ is versal if and only if $\mathcal{S} \oplus \mathcal{T} = \mathbf{R}^n$. Clearly $\mathcal{V}_1(\phi(q)) + \theta(q) \in \mathcal{S} \oplus \mathcal{T}$, and since $\mathcal{V}_2(p)$ may be arbitrary, it is necessary that $\text{span}(\{t_i\}) \oplus \mathcal{S} = \mathbf{R}^n$. It is also sufficient because we then obtain linear projections allowing us to write $\mathcal{V}_2(q) = x + \pi_{\mathcal{S}} \mathcal{V}_2(q) + \pi_{\mathcal{T}} \mathcal{V}_2(q)$. The functions ϕ and θ may be obtained from $\pi_{\mathcal{S}}$ and $\pi_{\mathcal{T}}$.

DEFINITION 2.3. *A linear deformation $\mathcal{V}(p)$ of the point x is* universal *or* miniversal *if it is versal and has the fewest possible parameters needed for a versal deformation.*

The number of parameters in a miniversal deformation is exactly the codimension of $\mathcal{S}$. Numerical analysts might prefer taking the t_i to be an orthogonal basis for $\mathcal{S}^{\perp}$, the subspace perpendicular to $\mathcal{S}$. This provides one natural miniversal deformation. Arnold [1] does not insist on using $\mathcal{S}^{\perp}$; any basis for any subspace of dimension $n - \dim \mathcal{S}$ will do provided that it intersects $\mathcal{S}$ at zero only. From the topological point of view, this is exactly the same, though of course the numerical properties may be quite different.

2.4. Versal deformations—the bigger picture. The previous section explained the linear or first-order theory of versal deformations. At this point, the reader might wonder whether this is just a whole lot of jargon to merely extend a basis for a subspace to the entire space. At the risk of delaying the motivation until now, we decided to make sure that the linear theory be well understood.

We are still in a finite-dimensional Euclidean space $\mathbf{R}^n$, but $\mathcal{S}$ will no longer be a flat subspace. Instead, we wish to consider any equivalence relation $\sim$ such that the *orbit* of x ($\text{orbit}(x) \equiv \{y | y \sim x\}$) is a sufficiently smooth submanifold. As an example we might define $x \sim y$ to mean $\|x\| = \|y\|$, in which case the orbits are spheres. In this context the word "orbit" is quite natural. In n^2-dimensional space points may be thought of as $n \times n$ matrices, and the orbit is the set of matrices with the same Jordan canonical form.

One final example that we must mention (because it explains the origins and significance of singularity theory) lives in an infinite-dimensional space. The vector space is the set of analytic functions $f(x)$ for which $f(0) = 0$. We can define $f \sim g$ if $f(x)$ and $g(\phi(x))$ have the same Taylor expansion at $x = 0$, where ϕ is a monotonic analytic function with $\phi(0) = 0$. The orbit of any function is some complicated infinite-dimensional manifold, but the codimension of the manifold happens to be finite.

Returning to $\mathbf{R}^n$, we can now cast everything into a nonlinear context.

DEFINITION 2.4. *A* deformation *of the point x is any continuously differentiable function*

$$\mathcal{V}(p_1, \ldots, p_l)$$

satisfying $\mathcal{V}(0) = x$.

DEFINITION 2.5. *A deformation* $\mathcal{V}_1(p)$ *of the point* x *is* versal *if for all deformations* $\mathcal{V}_2(q)$ *it is possible to write*

$$\mathcal{V}_2(q) \sim \mathcal{V}_1(\phi(q))$$

in an arbitrarily small neighborhood of 0, *where* $\phi(q)$ *is a continuously differentiable function from* $q_1, \ldots, q_m$ *to* $p_1, \ldots, p_l$ *for which* $\phi(0) = 0$.

The good news is that the inverse function theorem lets us express this nonlinear notion in terms of the linear theory.

THEOREM 2.6. *A deformation* $\mathcal{V}(p)$ *of* x *is* versal *if and only if* $\mathcal{V}_*(p)$ *is a versal linear deformation at the point* x *on the subspace* $\tan(\text{orbit}(x))$, *where* $\mathcal{V}_*(p)$ *is the linearization of* $\mathcal{V}(p)$ *near* x (*i.e., only first derivatives matter*) *and* tan *denotes the subspace tangent to the orbit at* x.

The rigorous proof may be found in [1], but the intuition should be clear: near the point x, only linear deformations matter, and the curvature of the orbit becomes unimportant—only the tangent plane matters. In other words, $y \sim x$ only if y is in the orbit of x, but to first order $y \sim x$ if (roughly speaking) $y = x + s$, where s is a small tangent vector to the orbit. The versality theorem (Theorem 2.6) shows that we only have to consider versal linear deformations, which we in the following denote $\mathcal{V}(p)$.

2.5. Versal deformations for the Jordan canonical form. We begin with deformations of the matrix $A = J_2(0)$. The perturbation theory and the normal and tangent spaces were discussed in section 1.3. We will use the same coordinate system here.

Four parameters $q = (q_1, q_2, q_3, q_4)$ are sufficient to describe the most general deformation of A:

$$\mathcal{V}_2(q) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} q_1 & q_2 \\ q_3 & q_4 \end{pmatrix}.$$

The equivalence relation is that of similar matrices, and it is easy to see by checking the trace and determinant that for sufficiently small values of q we have the equivalence

$$\mathcal{V}_2(q) \sim \mathcal{V}_1(p) \equiv \begin{pmatrix} 0 & 1 \\ p_1 & p_2 \end{pmatrix},$$

where $p = \phi(q)$ is defined by $p_1 = q_3(1 + q_2) - q_1 q_4$ and $p_2 = q_1 + q_4$. It is worth emphasizing that the equivalence relation does not work if $\mathcal{V}_2(q)$ is derogatory, but this does not happen for small parameters q.

We then see from Definition 2.5 that the two-parameter deformation $\mathcal{V}_1(p)$ is versal. In fact, it is *miniversal*, in that one needs the two parameters. From the local theory pictured in section 1.3, we saw that the orbit of $J_2(0)$ is the two-dimensional cone, and therefore the tangent and normal spaces are each two dimensional. The number of parameters in a miniversal deformation is always the dimension of the normal space.

It is a worthwhile exercise to derive the similarity transformation $S(q)$ (a deformation of the identity matrix) for which

$$\mathcal{V}_2(q) = S(q)^{-1} \mathcal{V}_1(\phi(q)) S(q),$$

and then linearize this map for small values of q to see which directions fall along the tangent space to the cone and which directions are normal to the cone.

Now consider deformations of $A = I_2$ or $A = 0$. Both matrices are derogatory with two eigenvalues 1 and 0, respectively. The tangent space does not exist (i.e., it is zero dimensional). Any possible behavior may be found near I_2 (or 0), including a one-dimensional space of derogatory matrices. The miniversal deformation of I_2 (or 0) is the full deformation requiring four parameters.

The general case has been worked out by Arnold [1]. The tangent vectors to the orbit of a matrix A are those matrices that may be expressed as $XA - AX$. The normal space is the adjoint of the centralizer, i.e., the set of matrices Z satisfying

$$A^H Z = Z A^H.$$

DEFINITION 2.7. *A deformation $\mathcal{V}(p) = A + Z(p)$ of a matrix A is a versal deformation if and only if $Z(p)$ is a basis for the orthogonal complement of* $\mathrm{orbit}(A)$ *that intersects the orbit at A.*

The formal definition of the similarity orbit of a matrix A is

$$\mathrm{orbit}(A) = \{S^{-1}AS : \det(S) \neq 0\}.$$

The parameterized normal form $Z(p)$ is the set of matrices that commute with A^H [1, 16]. For numerical properties we prefer taking $Z(p)$ to be an orthogonal basis for the normal space of $\mathrm{orbit}(A)$ at A. This choice of $Z(p)$ also ensures that $\mathcal{V}(p)$ is a miniversal deformation with one parameter for each dimension of the normal space.

Let A have r distinct eigenvalues $\lambda_i, i = 1 : r$ with r_i Jordan blocks each. Let $s_1(\lambda_i) \geq s_2(\lambda_i) \geq \cdots \geq s_{r_i}(\lambda_i)$ denote the sizes of the Jordan blocks corresponding to the eigenvalue λ_i. Then the dimension of the normal space of A is

$$\sum_{i=1}^{r}\sum_{j=1}^{r_i}(2j-1)s_j(\lambda_i) = \sum_{i=1}^{r}(s_1(\lambda_i) + 3s_2(\lambda_i) + 5s_3(\lambda_i) + \cdots).$$

Notice that the values of the distinct λ_i play no role in this formula. The dimension of the normal space of A is determined only by the sizes of the Jordan blocks of A associated with distinct eigenvalues. If the matrix is in Jordan canonical form, then the normal space consists of matrices $Z(p)$ made up of Toeplitz blocks, whose block structure is completely determined by the sizes of the Jordan blocks for different eigenvalues. The normal space is the same for all matrices with the same Jordan structure independent of the values of the distinct eigenvalues, so one may as well consider only Jordan blocks corresponding to a 0 eigenvalue. This form of the normal space for the 0 eigenvalues is a special case in Theorem 5.3.

3. The algebra of matrix pencils–canonical forms. We saw in section 2.4 that to consider versal deformations one needs a finite- or infinite-dimensional space and an equivalence relation on this space. For the remainder of this paper, we consider the finite-dimensional Euclidean space of matrix pencils endowed with the Euclidean metric (usually denoted the Frobenius metric in this context). The equivalence relation is that of the strict equivalence of pencils.

We consider a matrix pencil $A - \lambda B$, where A and B are arbitrary $m \times n$ matrices with real or complex entries. The pencil is said to be *regular* if $m = n$ and $\det(A - \lambda B)$ is not identically zero. Indeed, the zeros of $\det(A - \lambda B) = 0$ are the (generalized) eigenvalues of a regular pencil. Otherwise, i.e., if $\det(A - \lambda B)$ is identically zero or

$m \neq n$, $A - \lambda B$ is called *singular*. Two $m \times n$ pencils $A_1 - \lambda B_1$ and $A_2 - \lambda B_2$ are *strictly equivalent* if there exist constant (independent of λ) invertible matrices P of size $m \times m$ and Q of size $n \times n$ such that

$$P^{-1}(A_1 - \lambda B_1)Q = A_2 - \lambda B_2.$$

Kronecker has shown that any matrix pencil is strictly equivalent to a canonical diagonal form that describes the structure elements of $A - \lambda B$ (including generalized eigenvalues and eigenspaces) in full detail (e.g., see [16]). This form is a generalization of the Jordan canonical form (JCF) to general matrix pencils.

3.1. Kronecker canonical form. The *Kronecker canonical form* (KCF) of $A - \lambda B$ exhibits the fine structure elements, including elementary divisors (Jordan blocks) and minimal indices, and is defined as follows [16]. Suppose $A, B \in \mathbf{C}^{m\times n}$. Then there exist nonsingular $P \in \mathbf{C}^{m\times m}$ and $Q \in \mathbf{C}^{n\times n}$ such that

$$P^{-1}(A - \lambda B)Q = \tilde{A} - \lambda \tilde{B}, \tag{3.1}$$

where $\tilde{A} = \operatorname{diag}(A_1, \ldots, A_b)$ and $\tilde{B} = \operatorname{diag}(B_1, \ldots, B_b)$ are block diagonal. $A_i - \lambda B_i$ is $m_i \times n_i$. We can partition the columns of P and Q into blocks corresponding to the blocks of $\tilde{A} - \lambda\tilde{B}$: $P = [P_1, \ldots, P_b]$, where P_i is $m \times m_i$, and $Q = [Q_1, \ldots, Q_b]$, where Q_i is $n \times n_i$. Each block $M_i \equiv A_i - \lambda B_i$ must be of one of the following forms: $J_j(\alpha)$, N_j, L_j, or L_j^T. First we consider

$$J_j(\alpha) \equiv \begin{bmatrix} \alpha - \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \alpha - \lambda \end{bmatrix} \quad \text{and} \quad N_j \equiv \begin{bmatrix} 1 & -\lambda & & \\ & \ddots & \ddots & \\ & & \ddots & -\lambda \\ & & & 1 \end{bmatrix}.$$

$J_j(\alpha)$ is simply a $j \times j$ Jordan block, and α is called a *finite eigenvalue*. N_j is a $j \times j$ block corresponding to an *infinite eigenvalue* of multiplicity j. The $J_j(\alpha)$ and N_j blocks together constitute the *regular structure* of the pencil. All the $A_i - \lambda B_i$ are regular blocks if and only if $A - \lambda B$ is a regular pencil. $\sigma(A - \lambda B)$ denotes the eigenvalues of the regular part of $A - \lambda B$ (with multiplicities) and is called the *spectrum* of $A - \lambda B$.

The other two types of diagonal blocks are

$$L_j \equiv \begin{bmatrix} -\lambda & 1 & & \\ & \ddots & \ddots & \\ & & -\lambda & 1 \end{bmatrix} \quad \text{and} \quad L_j^T \equiv \begin{bmatrix} -\lambda & & \\ 1 & \ddots & \\ & \ddots & -\lambda \\ & & 1 \end{bmatrix}. \tag{3.2}$$

The $j \times (j+1)$ block L_j is called a *singular block of right* (*or column*) *minimal index* j. It has a one-dimensional right null space $[1, \lambda, \ldots, \lambda^j]^T$ for any λ. The $(j+1) \times j$ block L_j^T is a *singular block of left* (*or row*) *minimal index* j and has a one-dimensional left null space for any λ. The left and right singular blocks together constitute the *singular structure* of the pencil and appear in the KCF if and only if the pencil is singular. The regular and singular structures define the *Kronecker structure* of a singular pencil.

We also have a real KCF associated with real matrix pencils. If $A, B \in \mathbf{R}^{m\times n}$, there exist nonsingular $P \in \mathbf{R}^{m\times m}$ and $Q \in \mathbf{R}^{n\times n}$, where as before $P^{-1}(A - \lambda B)Q = \tilde{A} - \lambda\tilde{B}$ is block diagonal. The only difference with (3.1) is the Jordan blocks associated with complex conjugate pairs of eigenvalues. Let $\alpha = \mu + i\omega$, where μ, ω are real and

$\omega \neq 0$. If α is an eigenvalue of $A - \lambda B$, then $\bar{\alpha}$ is also an eigenvalue. Let $J_j(\alpha, \bar{\alpha})$ denote a Jordan block of size $2j \times 2j$ associated with a complex conjugate pair of eigenvalues, here illustrated with the case $j = 3$:

$$J_3(\alpha, \bar{\alpha}) \equiv \begin{bmatrix} \mu-\lambda & \omega & 1 & 0 & 0 & 0 \\ -\omega & \mu-\lambda & 0 & 1 & 0 & 0 \\ 0 & 0 & \mu-\lambda & \omega & 1 & 0 \\ 0 & 0 & -\omega & \mu-\lambda & 0 & 1 \\ 0 & 0 & 0 & 0 & \mu-\lambda & \omega \\ 0 & 0 & 0 & 0 & -\omega & \mu-\lambda \end{bmatrix}.$$

The Jordan block $J_j(\alpha, \bar{\alpha})$ plays the same role in the real JCF as $\mathrm{diag}(J_j(\alpha), J_j(\bar{\alpha}))$ does in the complex JCF. Notice that each pair of the $2j$ columns of the real P and Q associated with a $J_j(\alpha, \bar{\alpha})$ block form the real and imaginary parts of the (generalized) principal chains corresponding to the complex conjugate pair of eigenvalues.

3.2. Generalized Schur form and reducing subspaces. In most applications it is sufficient to transfer $A - \lambda B$ to a *generalized Schur form* (e.g., to GUPTRI form [11, 12])

$$P^H(A - \lambda B)Q = \begin{bmatrix} A_r - \lambda B_r & * & * \\ 0 & A_{reg} - \lambda B_{reg} & * \\ 0 & 0 & A_l - \lambda B_l \end{bmatrix}, \tag{3.3}$$

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

$A_r - \lambda B_r$ has only right minimal indices in its KCF, indeed the same L_j blocks as $A - \lambda B$. Similarly, $A_l - \lambda B_l$ has only left minimal indices in its KCF, the same L_j^T blocks as $A - \lambda B$. If $A - \lambda B$ is singular at least one of $A_r - \lambda B_r$ and $A_l - \lambda B_l$ will be present in (3.3). The explicit structure of the diagonal blocks in staircase form can be found in [12]. If $A - \lambda B$ is regular $A_r - \lambda B_r$ and $A_l - \lambda B_l$ are not present in (3.3) and the GUPTRI form reduces to the upper triangular block $A_{reg} - \lambda B_{reg}$. Staircase forms that reveal the Jordan structure of the zero and infinite eigenvalues are contained in $A_{reg} - \lambda B_{reg}$.

Given $A - \lambda B$ in GUPTRI form, we also know different pairs of reducing subspaces [33, 11]. Suppose the eigenvalues on the diagonal of $A_{reg} - \lambda B_{reg}$ are ordered so that the first k, say, are in Λ_1 (a subset of the spectrum) and the remainder are outside Λ_1. Let $A_r - \lambda B_r$ be $m_r \times n_r$. Then the left and right reducing subspaces corresponding to Λ_1 are spanned by the leading $m_r + k$ columns of P and leading $n_r + k$ columns of Q, respectively. When Λ_1 is empty, the corresponding reducing subspaces are called *minimal*, and when Λ_1 contains the whole spectrum the reducing subspaces are called *maximal*.

Several authors have proposed (staircase-type) algorithms for computing a generalized Schur form (e.g., see [2, 22, 24, 23, 31, 36]). They are numerically stable in the sense that they compute the exact Kronecker structure (generalized Schur form or something similar) of a nearby pencil $A' - \lambda B'$. $\delta \equiv \|(A - A', B - B')\|_E$ is an upper bound on the distance to the closest $(A + \delta A, B + \delta B)$ with the KCF of (A', B').

Recently, robust software with error bounds for computing the GUPTRI form of a singular $A - \lambda B$ has been published [11, 12]. Some computational experiments that use this software will be discussed later.

3.3. Generic and nongeneric Kronecker structures. Although the KCF looks quite complicated in the general case, most matrix pencils have a quite simple Kronecker structure. If $A - \lambda B$ is $m \times n$, where $m \neq n$, then for almost all A and B it will have the same KCF, depending only on m and n. This corresponds to the *generic case* when $A - \lambda B$ has full rank for any complex (or real) value of λ. Accordingly, generic rectangular pencils have no regular part. The generic Kronecker structure for $A - \lambda B$ with $d = n - m > 0$ is

$$\operatorname{diag}(L_\alpha, \ldots, L_\alpha, L_{\alpha+1}, \ldots, L_{\alpha+1}),$$

where $\alpha = \lfloor m/d \rfloor$, the total number of blocks is d, and the number of $L_{\alpha+1}$ blocks is $m \bmod d$ (which is 0 when d divides m) [31, 8]. The same statement holds for $d = m - n > 0$ if we replace $L_\alpha, L_{\alpha+1}$ in (3.2) by $L_\alpha^T, L_{\alpha+1}^T$. Square pencils are generically regular; i.e., $\det(A - \lambda B) = 0$ if and only if λ is an eigenvalue. The generic singular pencils of size $n \times n$ have the Kronecker structures [34]

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

Only if a singular $A - \lambda B$ is rank deficient (for some λ) may the associated KCF be more complicated and possibly include a regular part, as well as right and left singular blocks. This situation corresponds to the *nongeneric or degenerate case*, which of course is the real challenge from a computational point of view.

The generic and nongeneric cases can easily be couched in terms of reducing subspaces. For example, generic rectangular pencils have only trivial reducing subspaces and no generalized eigenvalues at all. Generic square singular pencils have the same minimal and maximal reducing subspaces. We think of a nongeneric case as an $A - \lambda B$ that lies either in a submanifold (its orbit) or the bundle corresponding to similar forms but with differing eigenvalues. In this case the pencil has nontrivial reducing subspaces. Moreover, only if it is perturbed so as to move continuously within this manifold or bundle does its reducing subspaces and generalized eigenvalues also move continuously and satisfy interesting error bounds [9, 11, 14, 26]. These requirements are natural in many control and systems theoretic problems, such as computing controllable subspaces and uncontrollable modes.

4. The geometry of matrix pencil space. In the coming sections we derive formulas for the tangent and normal spaces of the orbit of a matrix pencil that we will make use of in computing the versal form in section 5. We also derive new bounds for the distance to less generic pencils.

4.1. The orbit of a matrix pencil and its tangent and normal spaces. Any $m \times n$ matrix pair (A, B) (with real or complex entries) defines a manifold of *strictly equivalent* matrix pencils in the $2mn$-dimensional space $\mathcal{P}$ of $m \times n$ pencils:

$$\operatorname{orbit}(A - \lambda B) = \{P^{-1}(A - \lambda B)Q : \det(P)\det(Q) \neq 0\}. \tag{4.1}$$

We may choose a special element of $\operatorname{orbit}(A - \lambda B)$ that reveals the KCF of the pencil.

As usual, the dimension of $\operatorname{orbit}(A - \lambda B)$ is equal to the dimension of the tangent space to the orbit at $A - \lambda B$, here denoted $\tan(A - \lambda B)$. By considering the deformation $(I_m + \delta X)(A - \lambda B)(I_n - \delta Y)$ of $A - \lambda B$ to first-order term in δ, where δ is a

small scalar, we obtain $A-\lambda B+\delta(X(A-\lambda B)-(A-\lambda B)Y)+O(\delta^2)$, from which it is evident that $\tan(A-\lambda B)$ consists of the pencils that can be represented in the form

$$T_A-\lambda T_B=(XA-AY)-\lambda(XB-BY), \tag{4.2}$$

where X is an $m\times m$ matrix and Y is an $n\times n$ matrix. (This may also be obtained formally by differentiating the exponential map.)

In the language of pure mathematics the map that sends the triple $(P,Q,A-\lambda B)$ to $P^{-1}(A-\lambda B)Q$ is called a *group action*. The group is the ordered pair of nonsingular matrices (P,Q) denoted $GL_m\times GL_n$ which indicates the size of the matrices and the fact that they are nonsingular. The group $GL_m\times GL_n$ then is acting on the set of pencils.

A group action is *transitive* if it maps the set onto itself; i.e., if every member of the set may be reached from every other member of the set by the map. Clearly the group action is transitive on orbits. (This is merely a restatement of the definition of an orbit: an orbit is a minimal transitive set with respect to the group action.)

Since the action is transitive, we immediately have that orbits are manifolds. Intuitively, the tangent space "looks" the same at every point, since it may be moved from any point to another point by the group action. Mathematically, the orbit is a *homogeneous space*. The orbit may be equated with the quotient group obtained by forming equivalence classes of pairs (P,Q) that map $A-\lambda B$ to the same point. It is a small step to show that reducing subspaces vary smoothly if one perturbs a pencil so that it stays on the same orbit. All one must do is *lift* a curve (maintaining continuity) through a pencil back up to $GL_m\times GL_n$ and then project out the reducing subspaces.

Using Kronecker products, we can represent the $2mn$ vectors $T_A-\lambda T_B\in\tan(A-\lambda B)$ as

$$\begin{bmatrix}\operatorname{vec}(T_A)\\ \operatorname{vec}(T_B)\end{bmatrix}=\begin{bmatrix}A^T\otimes I_m\\ B^T\otimes I_m\end{bmatrix}\operatorname{vec}(X)-\begin{bmatrix}I_n\otimes A\\ I_n\otimes B\end{bmatrix}\operatorname{vec}(Y).$$

In this notation, we may say that the tangent space is the range of the $2mn\times(m^2+n^2)$ matrix

$$T\equiv\begin{bmatrix}A^T\otimes I_m & -I_n\otimes A\\ B^T\otimes I_m & -I_n\otimes B\end{bmatrix}. \tag{4.3}$$

We may define the *normal* space $\operatorname{nor}(A-\lambda B)$ as the space perpendicular to $\tan(A-\lambda B)$. Orthogonality in $\mathcal{P}$, the $2mn$-dimensional space of matrix pencils, is defined with respect to a Frobenius inner product

$$\langle A-\lambda B,C-\lambda D\rangle\equiv\operatorname{tr}(AC^H+BD^H),$$

where $\operatorname{tr}(X)$ denotes the trace of a square matrix X. Remembering that the space orthogonal to the range of a matrix is the kernel of the Hermitian transpose, we have that

$$\operatorname{nor}(A-\lambda B)=\ker(T^H)=\ker\begin{bmatrix}\bar{A}\otimes I_m & \bar{B}\otimes I_m\\ -I_n\otimes A^H & -I_n\otimes B^H\end{bmatrix}.$$

In ordinary matrix notation, this states that $Z_A-\lambda Z_B$ is in the normal space of $A-\lambda B$ if and only if

$$Z_AA^H+Z_BB^H=0\quad\text{and}\quad A^HZ_A+B^HZ_B=0. \tag{4.4}$$

The conditions on Z_A and Z_B can easily be verified and also be derived in terms of the Frobenius inner product, i.e.,

$$\langle T_A - \lambda T_B, Z_A - \lambda Z_B \rangle = \operatorname{tr}(X(AZ_A^H + BZ_B^H) - (Z_A^H A + Z_B^H B)Y). \tag{4.5}$$

Verification. If conditions (4.4) are satisfied, it follows from (4.5) that the inner product is zero.

Derivation. If $\langle T_A - \lambda T_B, Z_A - \lambda Z_B \rangle = 0$, then $\operatorname{tr}(X(AZ_A^H + BZ_B^H) - (Z_A^H A + Z_B^H B)Y) = 0$ must hold for any X (of size $m \times m$) and Y (of size $n \times n$). By choosing $X \equiv 0$, (4.5) reduces to $\operatorname{tr}((Z_A^H A + Z_B^H B)Y) = 0$, which holds for any Y if and only if $Z_A^H A + Z_B^H B = 0$. Similarly, we can choose $Y \equiv 0$, which gives that $AZ_A^H + BZ_B^H = 0$.

If $B = I$, this reduces to $Z_A \in \operatorname{nor}(A)$ if and only if $Z_A^H \in \operatorname{centralizer}(A)$, which is a well-known fact (e.g., see [1]). We will see in section 5.3 that though the A-part of the normal space is very simple when $B = I$, obtaining an orthonormal basis for the B-part is particularly challenging. The requirement that $Z_B = -A^H Z_A$ when $B = I$ destroys any orthogonality one may have in a basis for the A-part.

We now collect our general statements and a few obvious consequences.

THEOREM 4.1. *Let the $m \times n$ pencil $A - \lambda B$ be given. Define the $2mn \times (m^2 + n^2)$ matrix T as in* (4.3). *Then*

$$\tan(A - \lambda B) = \operatorname{range}(T) = \{(XA - AY) - \lambda(XB - BY)\},$$

where X and Y are compatible square matrices, and

$$\operatorname{nor}(A - \lambda B) = \ker(T^H) = \{Z_A - \lambda Z_B\},$$

where $Z_A A^H + Z_B B^H = 0$ and $A^H Z_A + B^H Z_B = 0$.

The dimensions of these spaces are

$$\dim(\tan(A - \lambda B)) = m^2 + n^2 - \dim(\ker(T)) \tag{4.6}$$

and

$$\dim(\operatorname{nor}(A - \lambda B)) = \dim(\ker(T^H)) = \dim(\ker(T)) - (m - n)^2. \tag{4.7}$$

Of course, the tangent and normal spaces are complementary and span the complete $2mn$-dimensional space, i.e., $\mathcal{P} = \tan(A - \lambda B) \oplus \operatorname{nor}(A - \lambda B)$, so that the dimensions in (4.6) and (4.7) add up to $2mn$, as they should.

Theorem 4.1 leads to one approach for computing a basis for $\operatorname{nor}(A - \lambda B)$ from the singular value decomposition (SVD) of T. Indeed, the left singular vectors corresponding to the zero singular value form such a basis. The dimension of the normal space is also known as the *codimension* of the orbit, here denoted $\operatorname{cod}(A - \lambda B)$. Accordingly, we have the following "compact" characterization of the codimension of $\operatorname{orbit}(A - \lambda B)$.

COROLLARY 4.2. *Let the $m \times n$ pencil $A - \lambda B$ be given. Then*

$$\operatorname{cod}(A - \lambda B) = \textit{the number of zero singular values of } T.$$

The corresponding result for the (square) matrix case is

$$\operatorname{cod}(A) = \text{the number of zero singular values of } I_n \otimes A - A^T \otimes I_n.$$

Although the SVD-based method is simple and has nice numerical properties (backward stability), it is rather costly in the number of operations. Computing the SVD of T is an $O(m^3n^3)$ operation.

Knowing the Kronecker structure of $A - \lambda B$, it is also possible to compute the codimension of the orbit as the sum of separate codimensions [8]:

$$\operatorname{cod}(A - \lambda B) = c_{\text{Jor}} + c_{\text{Right}} + c_{\text{Left}} + c_{\text{Jor,Sing}} + c_{\text{Sing}}. \tag{4.8}$$

The different contributions in (4.8) originate from the Jordan structure of all eigenvalues (including any infinite eigenvalue), the right singular blocks ($L_j \leftrightarrow L_k$), the left singular blocks ($L_j^T \leftrightarrow L_k^T$), interactions of the Jordan structure with the singular blocks (L_k and L_j^T), and interactions between the left and right singular structures ($L_j \leftrightarrow L_k^T$), respectively. Explicit expressions for these codimensions are derived in [8]. Assume that the given $A-\lambda B$ has $r \le \min(m,n)$ distinct eigenvalues $\lambda_i, i = 1:r$ with r_i Jordan blocks each. Let $s_1(\lambda_i) \ge s_2(\lambda_i) \ge \cdots \ge s_{r_i}(\lambda_i)$ denote the sizes of the Jordan blocks corresponding to the eigenvalue λ_i. Then the separate codimensions of (4.8) can be expressed as

$$c_{\text{Jor}} = \sum_{i=1}^{r}\sum_{j=1}^{r_i}(2j-1)s_j(\lambda_i) = \sum_{i=1}^{r}(s_1(\lambda_i) + 3s_2(\lambda_i) + 5s_3(\lambda_i) + \cdots),$$

$$c_{\text{Right}} = \sum_{j>k}(j-k-1), \quad c_{\text{Left}} = \sum_{j>k}(j-k-1), \quad c_{\text{Sing}} = \sum_{j,k}(j+k+2),$$

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

Notice that if we do not wish to specify the value of an eigenvalue λ_i, the codimension count for this unspecified eigenvalue is one less, i.e.,

$$-1 + s_1(\lambda_i) + 3s_2(\lambda_i) + 5s_3(\lambda_i) + \cdots.$$

This is sometimes done in algorithms for computing the Kronecker structure of a matrix pencil, where usually only the eigenvalues 0 and ∞ are specified and the remaining ones are unspecified.

It is possible to extract the Kronecker structure of $A - \lambda B$ from a generalized Schur decomposition in $O((\max(m,n))^3)$ operations. The most reliable SVD approach for computing a generalized Schur decomposition of $A - \lambda B$ requires at most $O((\max(m,n))^4)$ operations, which is still small compared to computing the SVD of T (4.3) for already moderate values of m and n (e.g., when $m = n$).

Speaking loosely, we refer to a pencil as having a particular codimension; when speaking strictly we mean that the orbit of the pencil has this codimension.

For given m and n the generic pencil has codimension 0 (i.e., spans the complete $2mn$-dimensional space), while the most nongeneric matrix pair $(A, B) = (0_{m\times n}, 0_{m\times n})$ has codimension $= 2mn$ (i.e., defines a "point" in $2mn$-dimensional space). Accordingly, any $m \times n$ nongeneric pencil different from the "zero pencil" has a codimension ≥ 1 and $< 2mn$.

4.2. A lower bound on the distance to a less generic pencil. The SVD characterization of the codimension of orbit$(A - \lambda B)$ in Corollary 4.2 leads to the following theorem, from which we present an interesting special case as a corollary.

THEOREM 4.3. *For a given $m \times n$ pencil $A - \lambda B$ with codimension c, a lower bound on the distance to the closest pencil $(A + \delta A) - \lambda(B + \delta B)$ with codimension $c + d$, where $d \geq 1$, is given by*

$$\|(\delta A, \delta B)\|_E \geq \frac{1}{\sqrt{m+n}} \left(\sum_{i=2mn-c-d+1}^{2mn} \sigma_i^2(T) \right)^{1/2}, \tag{4.9}$$

where $\sigma_i(T)$ denotes the ith largest singular value of T ($\sigma_i(T) \geq \sigma_{i+1}(T) \geq 0$).

Proof. It follows from Corollary 4.2 that T has rank $= 2mn - c$ if and only if $A - \lambda B$ has codimension c and $(A + \delta A) - \lambda(B + \delta B)$ has codimension $c + d$ ($d \geq 1$) if and only if $T + \delta T$, where δT is defined as

$$\delta T \equiv \begin{bmatrix} \delta A^T \otimes I_m & -I_n \otimes \delta A \\ \delta B^T \otimes I_m & -I_n \otimes \delta B \end{bmatrix}, \tag{4.10}$$

has rank $2mn - c - d$. From the construction, it follows that

$$\|\delta T\|_E = \sqrt{m+n}\|(\delta A, \delta B)\|_E$$

(each element δa_{ij} and δb_{ij} appears $m + n$ times in δT). The Eckart–Young and Mirsky theorem for finding the closest matrix of a given rank (e.g., see [17]) gives that the size of the smallest perturbation in Frobenius norm that reduces the rank in T from $2mn - c$ to $2mn - c - d$ is

$$\left(\sum_{i=2mn-c-d+1}^{2mn-c} \sigma_i^2(T) \right)^{1/2}. \tag{4.11}$$

Moreover, the fact that $A - \lambda B$ has codimension c implies that $\sigma_{2mn-c+1}(T) = \cdots = \sigma_{2mn}(T) = 0$. Since $\|\delta T\|_E$ must be larger than or equal to quantity (4.11), the proof is complete. ∎

COROLLARY 4.4. *For a given generic $m \times n$ pencil $A - \lambda B$, a lower bound on the distance to the closest nongeneric pencil $(A + \delta A) - \lambda(B + \delta B)$ is given by*

$$\|(\delta A, \delta B)\|_E \geq \frac{\sigma_{\min}(T)}{\sqrt{m+n}}, \tag{4.12}$$

where $\sigma_{\min}(T) = \sigma_{2mn}(T)$ denotes the smallest singular value of T, which is nonzero for a generic $A - \lambda B$.

We remark that the set of $m \times n$ matrix pencils does not include orbits of all codimensions from 1 to $2mn$.

One application of Corollary 4.4 is to characterize the distance to uncontrollability for a multiple input/multiple output linear system $E\dot{x}(t) = Fx(t) + Gu(t)$, where E and F are $p \times p$ matrices, G is $p \times q$ ($p \geq q$), and E is assumed to be nonsingular. If $A - \lambda B \equiv [G|F - \lambda E]$ is generic, the linear system is controllable (i.e., the dimension of the controllable subspace equals p) and a lower bound on the distance to the closest uncontrollable system is given by (4.12).

5. Versal deformations for the KCF. In the coming sections, we derive versal deformations which for us will mean the decomposition of arbitrary perturbations into the tangent and normal spaces of the orbits of equivalent pencils. Since the set of pencils is itself a vector space, the tangent and normal spaces to the orbits may be thought of as linear affine subplanes embedded in the space of pencils.

DEFINITION 5.1. *A deformation* $\mathcal{V}(p) = A - \lambda B + Z_A(p) - \lambda Z_B(p)$ *of a pencil* $A - \lambda B$ *is a versal deformation if and only if* $Z_A(p) - \lambda Z_B(p)$ *is a basis for the orthogonal complement of* $\mathrm{orbit}(A - \lambda B)$ *that intersects the orbit at* $A - \lambda B$.

Throughout this paper we will choose $Z_A(p) - \lambda Z_B(p)$ to have minimum number of parameters and to be an orthogonal basis for the normal space of $\mathrm{orbit}(A - \lambda B)$ at $A - \lambda B$. When it is clear from the context, we will drop the parameters and use the notation $Z_A - \lambda Z_B$ for the parameterized basis for the normal space.

5.1. An introductory example. We start with a small example before considering the general case. Let $A - \lambda B = L_1 \oplus L_4$ with codimension $= 2$. (This means that the manifold $\mathrm{orbit}(A - \lambda B)$ has codimension 2 or dimension 68 in the 70-dimensional space of 5×7 pencils.) Since $A - \lambda B$ is already in KCF we know its block structure:

$$A - \lambda B = \left[\begin{array}{cc|ccccc} -\lambda & 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & -\lambda & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\lambda & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\lambda & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\lambda & 1 \end{array}\right].$$

From (4.2) the matrices in the tangent space are given by $T_A - \lambda T_B = (XA - AY) - \lambda(XB - BY)$, where

$$T_A = \left[\begin{array}{cc|ccccc} -y_{21} & x_{11} - y_{22} & -y_{23} & x_{12} - y_{24} & x_{13} - y_{25} & x_{14} - y_{26} & x_{15} - y_{27} \\ \hline \boxed{-y_{41}} & x_{21} - y_{42} & -y_{43} & x_{22} - y_{44} & x_{23} - y_{45} & x_{24} - y_{46} & x_{25} - y_{47} \\ \boxed{-y_{51}} & \boxed{x_{31} - y_{52}} & -y_{53} & x_{32} - y_{54} & x_{33} - y_{55} & x_{34} - y_{56} & x_{35} - y_{57} \\ -y_{61} & \boxed{x_{41} - y_{62}} & -y_{63} & x_{42} - y_{64} & x_{43} - y_{65} & x_{44} - y_{66} & x_{45} - y_{67} \\ -y_{71} & x_{51} - y_{72} & -y_{73} & x_{52} - y_{74} & x_{53} - y_{75} & x_{54} - y_{76} & x_{55} - y_{77} \end{array}\right]$$

and

$$T_B = \left[\begin{array}{cc|ccccc} x_{11} - y_{11} & -y_{12} & x_{12} - y_{13} & x_{13} - y_{14} & x_{14} - y_{15} & x_{15} - y_{16} & -y_{17} \\ \hline x_{21} - y_{31} & -y_{32} & x_{22} - y_{33} & x_{23} - y_{34} & x_{24} - y_{35} & x_{25} - y_{36} & -y_{37} \\ \boxed{x_{31} - y_{41}} & -y_{42} & x_{32} - y_{43} & x_{33} - y_{44} & x_{34} - y_{45} & x_{35} - y_{46} & -y_{47} \\ \boxed{x_{41} - y_{51}} & \boxed{-y_{52}} & x_{42} - y_{53} & x_{43} - y_{54} & x_{44} - y_{55} & x_{45} - y_{56} & -y_{57} \\ x_{51} - y_{61} & \boxed{-y_{62}} & x_{52} - y_{63} & x_{53} - y_{64} & x_{54} - y_{65} & x_{55} - y_{66} & -y_{67} \end{array}\right].$$

By inspection we find the following two relations between elements in T_A and T_B:

$$\boxed{\square}: \; t^a_{21} + t^a_{32} = t^b_{31} + t^b_{42}$$

and

$$\square: \; t^a_{31} + t^a_{42} = t^b_{41} + t^b_{52},$$

where t^a_{ij} and t^b_{ij} denote the (i,j)th elements of T_A and T_B, respectively. These two relations clearly show that the tangent space has codimension at least 2. It may be verified that the other parameters may be chosen arbitrarily so that the codimension is exactly 2.

We want to find $Z_A - \lambda Z_B$ that is orthogonal to $T_A - \lambda T_B$ with respect to the Frobenius inner product, i.e.,

$$0 \equiv \langle T_A - \lambda T_B, Z_A - \lambda Z_B \rangle \equiv \operatorname{tr}(T_A Z_A^H + T_B Z_B^H) \equiv \sum_{i,j} t^a_{ij} \overline{z}^a_{ij} + t^b_{ij} \overline{z}^b_{ij}. \tag{5.1}$$

This inner product is most easily envisioned as the sum of the elementwise multiplication of the two pencils. Using this point of view, it is obvious that the normal space consists of pencils of the form $Z_A - \lambda Z_B \in \operatorname{nor}(A - \lambda B)$:

$$\begin{aligned} Z_A - \lambda Z_B &= \left[\begin{array}{cc|ccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline p_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_2 & p_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & p_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] - \lambda \left[\begin{array}{cc|ccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -p_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -p_2 & -p_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -p_2 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \\ &= \left[\begin{array}{cc|ccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline p_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_2 + \lambda p_1 & p_1 & 0 & 0 & 0 & 0 & 0 \\ \lambda p_2 & p_2 + \lambda p_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda p_2 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \end{aligned} \tag{5.2}$$

where p_1 and p_2 are arbitrary. Roughly speaking, the parameter p_1 corresponds to the doubly boxed entries ($\boxed{\square}$) and the parameter p_2 corresponds to the singly boxed entries. ($\square$).

Now, $\mathcal{V}(p) = A - \lambda B + Z_A - \lambda Z_B$ may be thought of as a versal deformation, or normal form, with minimum number of parameters (equal to the codimension of the original pencil). It follows that any (complex) pencil close to the given $A - \lambda B$ in KCF can be reduced to the two-parameter normal form $\mathcal{V}(p) = A - \lambda B + Z_A - \lambda Z_B$ in terms of equivalence transformations that are deformations of the identity.

5.2. Notation: A glossary of Toeplitz and Hankel matrices. The example in the previous section shows that a nonzero block of $Z_A - \lambda Z_B$ has a structured form. Indeed, the $(2,1)$ block has a Toeplitz-like form with $j - i = 3$ nonzero diagonals starting from the $(1,1)$ element of the $(2,1)$ block. A closer look shows that the A-part has $i - j - 1 = 2$ nonzero diagonals and the B-part is just the same matrix negated and with the diagonals shifted one row downward. In general, different nonzero blocks with Toeplitz or Hankel properties will show up in $Z_A - \lambda Z_B \in \operatorname{nor}(A - \lambda B)$. To simplify the proof of the general case we introduce some Toeplitz and Hankel matrices. Arrows and "stops" near the matrices make clear how the matrix is defined.

Let $S^L_{s\times t}$ be a *lower trapezoidal* $s \times t$ Toeplitz matrix with the first nonzero diagonal

starting at position $(1,1)$:

$$S^L_{s\times t} = \begin{matrix}\downarrow\\ \\ \\ \\ \\ \\ \perp\end{matrix}\!\!\begin{bmatrix} p_1 & 0 & 0 \\ \vdots & \ddots & 0 \\ \vdots & & p_1 \\ p_{s-t+1} & & \vdots \\ \vdots & \ddots & \vdots \\ p_s & \cdots & p_{s-t+1} \end{bmatrix} \text{ if } s \geq t \text{ and } S^L_{s\times t} = \begin{matrix}\downarrow\\ \\ \perp\end{matrix}\!\!\begin{bmatrix} p_1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ p_s & \cdots & p_1 & 0 \ \cdots 0 \end{bmatrix} \text{ otherwise,}$$

and let $T^L_{s\times t}$ be a *lower trapezoidal* $s \times t$ Toeplitz matrix with the first nonzero diagonal's last element at position (s,t):

$$T^L_{s\times t} = \begin{matrix}\\ \\ \\ \\ \\ \\ \vdash\end{matrix}\!\!\begin{bmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & & \vdots \\ p_1 & \ddots & \vdots \\ \vdots & \ddots & 0 \\ p_t & \cdots & p_1 \end{bmatrix}\!\!\begin{matrix}\\ \\ \\ \\ \\ \\ \leftarrow\end{matrix} \text{ if } s \geq t \text{ and } T^L_{s\times t} = \begin{matrix}\\ \vdash\end{matrix}\!\!\begin{bmatrix} p_{t-s+1} & \cdots & p_1 & 0 & 0 \\ \vdots & \ddots & & \ddots & 0 \\ p_t & \cdots & p_{t-s+1} & \cdots & p_1 \end{bmatrix}\!\!\begin{matrix}\\ \leftarrow\end{matrix} \text{ otherwise.}$$

If $s < t$, the entries of the last $t - s$ columns of $S^L_{s\times t}$ are zero. Similarly, if $s \geq t$, the entries of the first $s - t$ rows of $T^L_{s\times t}$ are zero.

Let $S^B_{s\times t}$ be a *banded lower trapezoidal* $s \times t$ Toeplitz with last row 0:

$$S^B_{s\times t} = \begin{matrix}\downarrow\\ \\ \\ \\ \perp\\ \\ \\ \end{matrix}\!\!\begin{bmatrix} p_1 & 0 & 0 \\ \vdots & \ddots & 0 \\ \vdots & & p_1 \\ p_{s-t} & & \vdots \\ 0 & \ddots & \vdots \\ \vdots & \ddots & p_{s-t} \\ 0 & \cdots & 0 \end{bmatrix} \text{ if } s > t \text{ and } S^B_{s\times t} = 0 \text{ otherwise,}$$

and let $T^B_{s\times t}$ be another *banded lower trapezoidal* $s \times t$ Toeplitz matrix, this time with last column 0:

$$T^B_{s\times t} = \begin{matrix}\vdash\\ \\ \\ \end{matrix}\!\!\begin{bmatrix} p_{t-s} & \cdots & p_1 & 0 & \cdots 0 \\ 0 & \ddots & & \ddots & \ddots \vdots \\ 0 & 0 & p_{t-s} & \cdots & p_1 \ 0 \end{bmatrix}\!\!\begin{matrix}\leftarrow\\ \\ \\ \end{matrix} \text{ if } s < t \text{ and } T^B_{s\times t} = 0 \text{ otherwise.}$$

Notice that the last row of $S^B_{s\times t}$ (if $s > t$) and the last column of $T^B_{s\times t}$ (if $s < t$) have all entries equal to zero.

Moreover, let $H^L_{s\times t}$ be a *lower trapezoidal* $s \times t$ Hankel matrix with the first nonzero diagonal starting at position $(1,t)$:

$$H^L_{s\times t} = \begin{bmatrix} 0 & 0 & p_1 \\ 0 & \iddots & \vdots \\ p_1 & & \vdots \\ \vdots & & p_{s-t+1} \\ \vdots & \iddots & \vdots \\ p_{s-t+1} & \cdots & p_s \end{bmatrix} \begin{matrix}\downarrow\\ \\ \\ \\ \\ \\ \bot\end{matrix} \text{ if } s \geq t \text{ and } H^L_{s\times t} = \begin{bmatrix} 0 & \cdots & 0 & p_1 \\ \vdots & & \iddots & \vdots \\ 0 \cdots & 0 & p_1 & p_s \end{bmatrix} \begin{matrix}\downarrow\\ \\ \bot\end{matrix} \text{ otherwise,}$$

and let $H^U_{s\times t}$ be a similar *upper trapezoidal* $s \times t$ Hankel matrix:

$$H^U_{s\times t} = \begin{bmatrix} \overset{\vdash}{p_t} & \cdots & \overset{\leftarrow}{p_1} \\ \vdots & \iddots & 0 \\ p_1 & \iddots & \vdots \\ 0 & & \vdots \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \text{ if } s \geq t \text{ and } H^U_{s\times t} = \begin{bmatrix} \overset{\vdash}{p_t} & \cdots & p_{t-s+1} & \cdots & \overset{\leftarrow}{p_1} \\ \vdots & \iddots & & \iddots & 0 \\ p_{t-s+1} & \cdots & p_1 & 0 & 0 \end{bmatrix} \text{ otherwise.}$$

If $s < t$, the entries of the first $t - s$ columns of $H^L_{s\times t}$ are zero. Similarly, if $s \geq t$, the entries of the last $s - t$ rows of $H^U_{s\times t}$ are zero.

Let $H_{s\times t}$ be a *dense* $s \times t$ Hankel matrix (with the first diagonal starting at position $(1,1)$):

$$H_{s\times t} = \begin{bmatrix} p_1 & p_2 & p_3 & \cdots & p_t \\ p_2 & \iddots & & & \vdots \\ p_3 & & & & \vdots \\ \vdots & & & & \vdots \\ p_s & & \cdots & & p_{s+t-1} \end{bmatrix}$$

for both the cases $s \geq t$ and $s < t$.

The *nilpotent* $k \times k$ matrix

$$C_k = \begin{bmatrix} 0 & I_{k-1} \\ 0 & 0 \end{bmatrix}$$

will be used as a shift operator. For a given $k \times n$ matrix X, the rows are shifted one row upward and downward by the operations $C_k X$ and $C_k^T X$, respectively. The columns are shifted one column rightward and one column leftward in an $n \times k$ matrix X by the operations XC_k and XC_k^T, respectively. The $k \times (k+1)$ matrices

$$G_k = [I_k \ 0] \text{ and } \hat{G}_k = [0 \ I_k],$$

will be used to pick all rows but one or all columns but one of a given matrix X in the following way. The first k and last k rows in a $(k+1) \times n$ matrix X are picked by

$G_k X$ and $\hat{G}_k X$, respectively. The k first and k last columns in an $n \times (k+1)$ matrix X are picked by XG_k^T and $X\hat{G}_k^T$, respectively.

Let $\hat{I}_k$ denote the $k \times k$ matrix obtained by reversing the order of the columns in the $k \times k$ identity matrix. It follows that for an $n \times k$ matrix X, the order of the columns is reversed by the multiplication $X\hat{I}_k$.

So far, the matrices introduced are rectangular Toeplitz and Hankel matrices with a special structure, e.g., lower trapezoidal (S^L, T^L, H^L), banded lower trapezoidal (S^B, T^B), upper trapezoidal (H^U), or dense (H). The matrices C and $G, \hat{G}$ that will be used as "shift" and "pick" operators, respectively, are Toeplitz matrices with only one nonzero diagonal. In the next section we will see that versal deformations for all combinations of different blocks in the KCF, except Jordan blocks with nonzero finite eigenvalues, can be expressed in terms of these matrices. To cope with nonzero finite Jordan blocks $J_k(\gamma), \gamma \neq 0$ we need to introduce three more matrices. First, we introduce two lower triangular Toeplitz matrices D^L and E^L, which are involved in the case with two $J_k(\gamma)$ blocks. Finally, we introduce the "monstrous" matrix F^D, which captures the cases with a (left or right) singular block and a $J_k(\gamma)$ block.

Given $\gamma \neq \{0, \infty\}$, define two infinite sequences of numbers d_i and e_i by the recursion

$$\begin{bmatrix} d_i \\ \gamma e_i \end{bmatrix} = - \begin{bmatrix} 1 & 1 \\ 1 & 2 - 1/i \end{bmatrix} \begin{bmatrix} \overline{\gamma} d_{i-1} \\ e_{i-1} \end{bmatrix} \tag{5.3}$$

starting with

$$\begin{bmatrix} d_1 \\ e_1 \end{bmatrix} = \begin{bmatrix} \gamma \\ 1 \end{bmatrix}.$$

Given sizes s and t, for $1 \leq q \leq \min\{s,t\}$, we define $D_{s\times t}[q]$ and $E_{s\times t}[q]$ as *lower triangular* Toeplitz matrices with q diagonals in terms of $d_1, \ldots, d_q$ and $e_1, \ldots, e_{q-1}$ and a *boundary* value $e_q^* = -\overline{\gamma} d_q$:

$$D_{s\times t}[q] = \begin{bmatrix} 0 & \cdots & & 0 \\ d_q & & & \\ d_{q-1} & \ddots & & \\ \vdots & \ddots & \ddots & \vdots \\ d_2 & & \ddots & \ddots \\ d_1 & d_2 & \cdots & d_{q-1} \; d_q \; 0 \end{bmatrix} \quad \text{and} \quad E_{s\times t}[q] = \begin{bmatrix} 0 & \cdots & & 0 \\ e_q^* & & & \\ e_{q-1} & \ddots & & \\ \vdots & \ddots & \ddots & \vdots \\ e_2 & & \ddots & \ddots \\ e_1 & e_2 & \cdots & e_{q-1} \; e_q^* \; 0 \end{bmatrix}.$$

We take linear combinations with parameters p_j to form the matrices

$$D_{s\times t}^L = \sum_{i=1}^{\min\{s,t\}} p_j D_{s\times t}[i]\,\pi(i) \quad \text{and} \quad E_{s\times t}^L = \sum_{i=1}^{\min\{s,t\}} p_j E_{s\times t}[i]\,\pi(i), \tag{5.4}$$

where $j = \min\{s,t\} - i + 1$ and $\pi(i) = -\prod_{k=2}^{i-1} k\gamma/(1-2k)$ is defined to be $1/\gamma$ and -1 for $i = 1$ and $i = 2$, respectively. The parameter index j and the scaling function $\pi(i)$ are chosen to satisfy $D_{s\times t}^L = S_{s\times t}^L$ and $E_{s\times t}^L = -C_s^T S_{s\times t}^L$ for $\gamma = 0$ in Theorem 5.3 (see Tables 5.1 and 5.2). By simplifying (5.4) using $i = j$ and $\pi(i) = 1$, this consistency will be lost, but we will still have valid expressions for the versal deformations.

The relations between the elements of $D^L_{s\times t}$ and $E^L_{s\times t}$ are most readily shown by an example:

$$D^L_{4,3} = \begin{bmatrix} 0 & 0 & 0 \\ p_1\left(\frac{2|\gamma|^4}{3}+\frac{4|\gamma|^2}{3}+1\right) & 0 & 0 \\ p_1\left(-\frac{2\gamma|\gamma|^2}{3}-\frac{2\gamma}{3}\right)+p_2\left(|\gamma|^2+1\right) & p_1\left(\frac{2|\gamma|^4}{3}+\frac{4|\gamma|^2}{3}+1\right) & 0 \\ p_1\frac{2\gamma^2}{3}-p_2\gamma+p_3 & p_1\left(-\frac{2\gamma|\gamma|^2}{3}-\frac{2\gamma}{3}\right)+p_2\left(|\gamma|^2+1\right) & p_1\left(\frac{2|\gamma|^4}{3}+\frac{4|\gamma|^2}{3}+1\right) \end{bmatrix}$$

and

$$E^L_{4,3} = \begin{bmatrix} 0 & 0 & 0 \\ p_1\left(-\frac{2\overline{\gamma}|\gamma|^4}{3}-\frac{4\overline{\gamma}|\gamma|^2}{3}-\overline{\gamma}\right) & 0 & 0 \\ p_1\left(-\frac{2|\gamma|^2}{3}-1\right)+p_2\left(-\overline{\gamma}|\gamma|^2-\overline{\gamma}\right) & p_1\left(-\frac{2\overline{\gamma}|\gamma|^4}{3}-\frac{4\overline{\gamma}|\gamma|^2}{3}-\overline{\gamma}\right) & 0 \\ p_1\frac{2\gamma}{3}-p_2-p_3\overline{\gamma} & p_1\left(-\frac{2|\gamma|^2}{3}-1\right)+p_2\left(-\overline{\gamma}|\gamma|^2-\overline{\gamma}\right) & p_1\left(-\frac{2\overline{\gamma}|\gamma|^4}{3}-\frac{4\overline{\gamma}|\gamma|^2}{3}-\overline{\gamma}\right) \end{bmatrix}.$$

Let $F^D_{s\times t}$ (D for dense) be defined as

$$F^D_{s\times t} = \sum_{i=1}^{s} p_{s-i+1}\ F_{s\times t}[i],$$

where $F_{s\times t}[q]$ has the q last rows nonzero and defined as

$$\begin{array}{ll} f_{s-q+1,j} = \overline{\gamma}^{j-1} & \text{for } j = 1,\dots,t, \\ f_{i,j} = \overline{\gamma} f_{i,j-1} + f_{i-1,j-1} & \text{for } i = s-q+2,\dots,s,\ j = 2,\dots,t, \end{array} \tag{5.5}$$

and $f_{i,1}$ for $i = s-q+2,\dots,s$ is defined as the solution to

$$\langle F_{s\times t}[q]G^T_{t-1} - \lambda F_{s\times t}[q]\hat{G}^T_{t-1},\ F_{s\times t}[s-i+1]G^T_{t-1} - \lambda F_{s\times t}[s-i+1]\hat{G}^T_{t-1}\rangle \equiv 0.$$

Notice that $f_{i,1}$ is used as an unknown in the generation of elements in (5.5). In the definition of $F_{s\times t}[q]$, the solutions for $f_{i,1}$ for $i = s-q+2,\dots,s$ ensure that $F_{s\times t}[q]G^T_{t-1} - \lambda F_{s\times t}[q]\hat{G}^T_{t-1}$ is orthogonal to $F_{s\times t}[\hat{q}]G^T_{t-1} - \lambda F_{s\times t}[\hat{q}]\hat{G}^T_{t-1}$ for $\hat{q} = 1,\dots,q-1$.

Also here we show a small example to facilitate the interpretation of the definition:

$$F^D_{3\times 2} = \begin{bmatrix} p_1 & p_1\overline{\gamma} \\ p_2 - p_1\frac{(|\gamma|^2+1)\gamma}{|\gamma|^4+2|\gamma|^2+2} & p_2\overline{\gamma} + p_1\frac{|\gamma|^2+2}{|\gamma|^4+2|\gamma|^2+2} \\ p_3 - p_2\frac{\gamma}{|\gamma|^2+1} + p_1\frac{\gamma^2}{|\gamma|^4+2|\gamma|^2+2} & p_3\overline{\gamma} + p_2\frac{1}{|\gamma|^2+1} - p_1\frac{\gamma}{|\gamma|^4+2|\gamma|^2+2} \end{bmatrix}.$$

5.3. Versal deformations—the general case. Without loss of generality assume that $A-\lambda B$ is already in KCF, $M = \operatorname{diag}(M_1, M_2, \dots, M_b)$, where each M_k is either a Jordan block associated with a finite or infinite eigenvalue or a singular block corresponding to a left or right minimal index. A pencil $T_A - \lambda T_B = XM - MY$ in the tangent space can be partitioned conformally with the pencil M so that

$T_{ij}^A - \lambda T_{ij}^B = X_{ij}M_j - M_iY_{ij}$, where M_k is $m_k \times n_k$, X_{ij} is $m_i \times m_j$, and Y_{ij} is $n_i \times n_j$:

$$\begin{bmatrix} X_{11} & \cdots & X_{1b} \\ \vdots & \ddots & \vdots \\ X_{b1} & \cdots & X_{bb} \end{bmatrix} \begin{bmatrix} M_1 & & \\ & \ddots & \\ & & M_b \end{bmatrix} - \begin{bmatrix} M_1 & & \\ & \ddots & \\ & & M_b \end{bmatrix} \begin{bmatrix} Y_{11} & \cdots & Y_{1b} \\ \vdots & \ddots & \vdots \\ Y_{b1} & \cdots & Y_{bb} \end{bmatrix}.$$

Since the blocks $T_{ij}^A - \lambda T_{ij}^B, i, j = 1, \ldots, b$ are mutually independent, we can study the different blocks of $T_A - \lambda T_B$ separately. Let $Z_{ij}^A - \lambda Z_{ij}^B$ be conformally sized blocks of $Z_A - \lambda Z_B$. From (4.4) we know that $Z_A - \lambda Z_B$ is in the normal space if and only if $A^H Z_A + B^H Z_B = 0$ and $Z_A A^H + Z_B B^H = 0$. We obtain a simple result since A and B are block diagonal.

PROPOSITION 5.2. *Assume that*

$$M = A - \lambda B = \mathrm{diag}(A_1, A_2, \ldots, A_b) - \lambda \mathrm{diag}(B_1, B_2, \ldots, B_b)$$

is in KCF, where each block $A_i - \lambda B_i \equiv M_i$ represents one block in the Kronecker structure. Then $Z_A - \lambda Z_B \in \mathrm{nor}(A - \lambda B)$ if and only if

$$A_j^H Z_{ji}^A = -B_j^H Z_{ji}^B \quad \text{and} \quad Z_{ji}^A A_i^H = -Z_{ji}^B B_i^H \quad \text{for } i = 1, \ldots, b \text{ and } j = 1, \ldots, b.$$

The mutual independency of the (i, j) blocks of Z_A and Z_B implies that we only have to consider two M_k blocks at a time:

$$\begin{aligned} T_A[i,j] - \lambda T_B[i,j] &= \begin{bmatrix} X_{ii} & X_{ij} \\ X_{ji} & X_{jj} \end{bmatrix} \begin{bmatrix} M_i & 0 \\ 0 & M_j \end{bmatrix} - \begin{bmatrix} M_i & 0 \\ 0 & M_j \end{bmatrix} \begin{bmatrix} Y_{ii} & Y_{ij} \\ Y_{ji} & Y_{jj} \end{bmatrix} \\ &= \begin{bmatrix} T_{ii}^A & T_{ij}^A \\ T_{ji}^A & T_{jj}^A \end{bmatrix} - \lambda \begin{bmatrix} T_{ii}^B & T_{ij}^B \\ T_{ji}^B & T_{jj}^B \end{bmatrix} \end{aligned}$$

and

$$Z_A[i,j] - \lambda Z_B[i,j] = \begin{bmatrix} Z_{ii}^A & Z_{ij}^A \\ Z_{ji}^A & Z_{jj}^A \end{bmatrix} - \lambda \begin{bmatrix} Z_{ii}^B & Z_{ij}^B \\ Z_{ji}^B & Z_{jj}^B \end{bmatrix}. \tag{5.6}$$

Notably, by interchanging the blocks $M_i = A_i - \lambda B_i$ and $M_j = A_j - \lambda B_j$ in the KCF, we only have to interchange the corresponding blocks in $Z_A - \lambda Z_B$ accordingly. For example, if $Z_A[i,j] - \lambda Z_B[i,j]$ in (5.6) belongs to $\mathrm{nor}(\mathrm{diag}(M_i, M_j))$, then

$$\begin{bmatrix} Z_{jj}^A & Z_{ji}^A \\ Z_{ij}^A & Z_{ii}^A \end{bmatrix} - \lambda \begin{bmatrix} Z_{jj}^B & Z_{ji}^B \\ Z_{ij}^B & Z_{ii}^B \end{bmatrix} \in \mathrm{nor}(\mathrm{diag}(M_j, M_i)).$$

This implies that given two blocks M_i and M_j, it is sufficient to consider the case $\mathrm{diag}(M_i, M_j)$. In the following we will order the blocks in the KCF so that $Z_A - \lambda Z_B$ is block lower triangular.

THEOREM 5.3. *Let $A - \lambda B = \mathrm{diag}(A_1, A_2, \ldots, A_b) - \lambda \mathrm{diag}(B_1, B_2, \ldots, B_b)$ be in KCF with the structure blocks $M_i = A_i - \lambda B_i$ ordered as follows: L_k, $J_k(0)$, $J_k(\gamma)$ (for $\gamma \neq \{0, \infty\}$), N_k, and L_k^T, where the ordering within each block type is in increasing order of size, except for the L_k^T blocks, which are ordered by decreasing order of size.*

For all i and j, let the $(i,j),(j,i)$ and $(i,i),(j,j)$ blocks of $Z_A(p) - \lambda Z_B(p)$ corresponding to $\mathrm{diag}(M_i, M_j)$ be built from Table 5.1 and Table 5.2, respectively.

TABLE 5.1

Blocks in $Z_A - \lambda Z_B \in \mathrm{nor}(A - \lambda B)$, where for $L_\alpha \oplus L_\beta$, $J_\alpha(0) \oplus J_\beta(0)$, $J_\alpha(\gamma) \oplus J_\beta(\gamma)$, and $N_\alpha \oplus N_\beta$ it is assumed that $\alpha \le \beta$. For $L_\alpha^T \oplus L_\beta^T$, $\alpha \ge \beta$ is assumed. Also $\gamma_1 \ne \gamma_2$ is assumed.

KCF:$M_i \oplus M_j$	Z_{ij}^A	Z_{ij}^B	Z_{ji}^A	Z_{ji}^B
$L_\alpha \oplus L_\beta$	0	0	$S^B_{\beta\times(\alpha+1)}$	$-C_\beta^T S^B_{\beta\times(\alpha+1)}$
$L_\alpha \oplus J_\beta(0)$	0	0	$S^L_{\beta\times(\alpha+1)}$	$-C_\beta^T S^L_{\beta\times(\alpha+1)}$
$L_\alpha \oplus J_\beta(\gamma)$	0	0	$F^D_{\beta\times(\alpha+2)} G^T_{\alpha+1}$	$-F^D_{\beta\times(\alpha+2)} \hat{G}^T_{\alpha+1}$
$L_\alpha \oplus N_\beta$	0	0	$C_\beta^T H^L_{\beta\times(\alpha+1)}$	$-H^L_{\beta\times(\alpha+1)}$
$L_\alpha \oplus L_\beta^T$	0	0	$G_{\beta+1} H_{(\beta+2)\times(\alpha+1)}$	$-\hat{G}_{\beta+1} H_{(\beta+2)\times(\alpha+1)}$
$J_\alpha(0) \oplus J_\beta(0)$	$S^L_{\alpha\times\beta}$	$-C_\alpha^T S^L_{\alpha\times\beta}$	$T^L_{\beta\times\alpha}$	$-C_\beta^T T^L_{\beta\times\alpha}$
$J_\alpha(0) \oplus L_\beta^T$	0	0	$H^U_{(\beta+1)\times\alpha}$	$-H^U_{(\beta+1)\times\alpha} C_\alpha^T$
$J_\alpha(\gamma) \oplus J_\beta(\gamma)$	$D^L_{\alpha\times\beta}$	$E^L_{\alpha\times\beta}$	$D^L_{\beta\times\alpha}$	$E^L_{\beta\times\alpha}$
$J_\alpha(\gamma) \oplus L_\beta^T$	0	0	$G_{\beta+1}(\hat{I}_\alpha F^D_{\alpha\times(\beta+2)})^T$	$-\hat{G}_{\beta+1}(\hat{I}_\alpha F^D_{\alpha\times(\beta+2)})^T$
$N_\alpha \oplus N_\beta$	$C_\alpha^T S^L_{\alpha\times\beta}$	$-S^L_{\alpha\times\beta}$	$C_\beta^T T^L_{\beta\times\alpha}$	$-T^L_{\beta\times\alpha}$
$N_\alpha \oplus L_\beta^T$	0	0	$T^L_{(\beta+1)\times\alpha} C_\alpha^T$	$-T^L_{(\beta+1)\times\alpha}$
$L_\alpha^T \oplus L_\beta^T$	0	0	$T^B_{(\beta+1)\times\alpha}$	$-T^B_{(\beta+1)\times\alpha} C_\alpha$
$J_\alpha(0) \oplus J_\beta(\gamma)$	0	0	0	0
$J_\alpha(0) \oplus N_\beta$	0	0	0	0
$J_\alpha(\gamma_1) \oplus J_\beta(\gamma_2)$	0	0	0	0
$J_\alpha(\gamma) \oplus N_\beta$	0	0	0	0

TABLE 5.2

The diagonal blocks in $Z_A - \lambda Z_B \in \mathrm{nor}(A - \lambda B)$.

KCF:M_i	Z_{ii}^A	Z_{ii}^B
L_α	0	0
$J_\alpha(0)$	$S^L_{\alpha\times\alpha}$	$-C_\alpha^T S^L_{\alpha\times\alpha}$
$J_\alpha(\gamma)$	$D^L_{\alpha\times\alpha}$	$E^L_{\alpha\times\alpha}$
N_α	$C_\alpha^T S^L_{\alpha\times\alpha}$	$-S^L_{\alpha\times\alpha}$
L_α^T	0	0

Then $Z_A(p) - \lambda Z_B(p)$ gives an orthogonal basis for $\mathrm{nor}(A - \lambda B)$ with minimum number of parameters; i.e., $\mathcal{V}(p) = A - \lambda B + Z_A(p) - \lambda Z_B(p)$ is a miniversal deformation of $A - \lambda B$.

The superscripts B, L, U, and D of the matrices in Tables 5.1 and 5.2 are parts of the matrix definitions in section 5.2. The superscript T is the matrix transpose. All subscripts, e.g., $\alpha \times \beta$, refer to the sizes of the matrices.

Notice that the diagonal blocks (i, i) and (j, j) of $Z_A - \lambda Z_B$ can also be obtained from Table 5.1 by setting $i = j$. For clarity we also display the expressions for the (i, i) and (j, j) blocks of $Z_A - \lambda Z_B$ corresponding to all kinds of structure blocks M_i in Table 5.2. Of course, the (j, j) blocks corresponding to M_j are read from Table 5.2 by substituting α with β.

The proof of Theorem 5.3 consists of three parts.

1. The blocks of $Z_A - \lambda Z_B$ displayed in Table 5.1 fulfill the conditions in Proposition 5.2, which imply that $Z_A - \lambda Z_B \in \mathrm{nor}(A - \lambda B)$ is orthogonal to an arbitrary $T_A - \lambda T_B \in \tan(A - \lambda B)$.
2. The number of independent parameters in $Z_A - \lambda Z_B$ is equal to the codimension of $\mathrm{orbit}(A - \lambda B)$, which implies that the parameterized normal form has minimum number of parameters.
3. Each block in Table 5.1 defines an orthogonal basis; i.e., the basis for each parameter p_i is orthogonal to the basis for each other parameter p_j, $i \neq j$.

We start by proving part 3 and then prove parts 1 and 2 for the 16 different cases $\mathrm{diag}(M_i, M_j)$ corresponding to different combinations of structure blocks in the KCF. In Table 5.3 we display the codimension for these 16 cases and the number of parameters in the $(i,i), (i,j), (j,i)$, and (j,j) blocks of $Z_A - \lambda Z_B$. The codimensions are computed from (4.8), which is the minimum number of parameters required to span the corresponding normal space. For the ordering and the sizes of the blocks in $A - \lambda B$ we have made the same assumptions in Table 5.3 as in Table 5.1. Notice that the codimension counts for $L_\alpha \oplus L_\beta$ and $L_\alpha^T \oplus L_\beta^T$ are 0 if $\alpha = \beta$. The number of parameters required in each of the $(i,i), (i,j), (j,i)$, and (j,j) blocks of $Z_A - \lambda Z_B$ follows from the proof given below.

TABLE 5.3

The number of parameters in the (i,i), (i,j), (j,i), and (j,j) blocks of $Z_A - \lambda Z_B \in \mathrm{nor}(M_i \oplus M_j)$.

KCF: $M_i \oplus M_j$	$\mathrm{cod}(M_i \oplus M_j)$	(i,i)	(i,j)	(j,i)	(j,j)
$L_\alpha \oplus L_\beta$	$\beta - \alpha - 1$	0	0	$\beta - \alpha - 1$	0
$L_\alpha \oplus J_\beta(0)$	2β	0	0	β	β
$L_\alpha \oplus J_\beta(\gamma)$	2β	0	0	β	β
$L_\alpha \oplus N_\beta$	2β	0	0	β	β
$L_\alpha \oplus L_\beta^T$	$\alpha + \beta + 2$	0	0	$\alpha + \beta + 2$	0
$J_\alpha(0) \oplus J_\beta(0)$	$\beta + 3\alpha$	α	α	α	β
$J_\alpha(0) \oplus L_\beta^T$	2α	α	0	α	0
$J_\alpha(\gamma) \oplus J_\beta(\gamma)$	$\beta + 3\alpha$	α	α	α	β
$J_\alpha(\gamma) \oplus L_\beta^T$	2α	α	0	α	0
$N_\alpha \oplus N_\beta$	$\beta + 3\alpha$	α	α	α	β
$N_\alpha \oplus L_\beta^T$	2α	α	0	α	0
$L_\alpha^T \oplus L_\beta^T$	$\alpha - \beta - 1$	0	0	$\alpha - \beta - 1$	0
$J_\alpha(0) \oplus J_\beta(\gamma)$	$\alpha + \beta$	α	0	0	β
$J_\alpha(0) \oplus N_\beta$	$\alpha + \beta$	α	0	0	β
$J_\alpha(\gamma_1) \oplus J_\beta(\gamma_2)$	$\alpha + \beta$	α	0	0	β
$J_\alpha(\gamma) \oplus N_\beta$	$\alpha + \beta$	α	0	0	β

To fully appreciate this rather technical proof it could be more fruitful to look first at some examples of versal deformations in section 6.1.

Proof of part 3. We show that each matrix pencil block in Table 5.1 has all its parameters in orthogonal directions. This is trivial for blocks built from the structured Toeplitz and Hankel matrices S^L, S^B, H, H^L, H^U, T^L, or T^B (possibly involving some kind of shift). Remember that the Frobenius inner product can be expressed in terms of the sum of all results from elementwise multiplications as shown in (5.1). For each of these matrices, the elementwise multiplication of the basis for one parameter p_i and the basis for another parameter $p_j, j \neq i$ only results in multiplications where at least one of the two elements is zero. Obviously, these bases are orthogonal. For the matrix pencil blocks built from the F^D matrix, the orthogonality follows from

construction since some of the elements are explicitly chosen so that the Frobenius inner product is zero.

For the proof for the blocks of type $D^L - \lambda E^L$ we define s_q in terms of the d_i and e_i in (5.3) to be

$$s_q = \sum_{i=1}^{q} i|d_i|^2 + \sum_{i=1}^{q-1} i|e_i|^2 - q\overline{\gamma} d_q \overline{e}_q.$$

Independent of s and t, the number s_q is the inner product of the qth basis vector with the rth, where $q < r$.

We show by induction that $s_q = 0$ for $q = 1, 2, \ldots$. Clearly $s_1 = |\gamma|^2 - \gamma\overline{\gamma} = 0$.

We now show that $s_{q+1} - s_q = 0$, from which the result follows:

$$\begin{aligned}
&q\overline{\gamma} d_q \overline{e}_q + (q+1)|d_{q+1}|^2 + q|e_q|^2 - (q+1)\overline{\gamma} d_{q+1}\overline{e}_{q+1} \\
&= q\overline{e}_q(\overline{\gamma} d_q + e_q) + (q+1)d_{q+1}(\overline{d}_{q+1} - \overline{\gamma}\ \overline{e}_{q+1}) \\
&= d_{q+1}((q+1)(\overline{d}_{q+1} - \overline{\gamma}\ \overline{e}_{q+1}) - q\overline{e}_q) \\
&= d_{q+1}\left((q+1)\left(-\gamma\overline{d}_q - \overline{e}_q + \gamma\overline{d}_q + 2\overline{e}_q - \frac{\overline{e}_q}{q+1}\right) - q\overline{e}_q\right) \\
&= d_{q+1}((q+1)\overline{e}_q - \overline{e}_q - q\overline{e}_q) = 0.
\end{aligned}$$

Since $Z_A - \lambda Z_B$ is built from b^2 mutually independent blocks in Table 5.1, each associated with c_i parameters, it follows that $Z_A - \lambda Z_B$ is an orthogonal basis for a $(c_1 + c_2 + \cdots + c_{b^2})$-dimensional space, with one parameter for each dimension. ∎

Proof of parts 1 *and* 2. Now, it remains to show that $Z_A - \lambda Z_B$ is orthogonal to $\tan(A - \lambda B)$ and that the number of parameters in $Z_A - \lambda Z_B$ is equal to $\operatorname{cod}(A - \lambda B)$. Since the number of parameters in orthogonal directions cannot exceed the codimension, it is sufficient to show that we have found them all. The orthogonality between $Z_A - \lambda Z_B$ and $\tan(A - \lambda B)$ is shown by proving that each pair of blocks fulfills the conditions $A_j^H Z_{ji}^A = -B_j^H Z_{ji}^B$ and $Z_{ji}^A A_i^H = -Z_{ji}^B B_i^H$ in Proposition 5.2. In the following we refer to these as the *first* and *second* conditions, respectively.

We carry out the proofs for all 16 cases $M_i \oplus M_j$ in Table 5.1, starting with blocks where M_i and M_j are of the same kind.

$\mathbf{J}_\alpha(\mathbf{0}) \oplus \mathbf{J}_\beta(\mathbf{0})$: We note that $J_k(0) = C_k - \lambda I_k$. *First* condition for the (j, i) block:

$$A_j^H Z_{ji}^A = C_\beta^T T_{\beta\times\alpha}^L = I_\beta C_\beta^T T_{\beta\times\alpha}^L = -B_j^H Z_{ji}^B.$$

Second condition for the (j, i) block:

$$Z_{ji}^A A_i^H = T_{\beta\times\alpha}^L C_\alpha^T = T_{\beta\times\alpha}^L C_\alpha^T I_\alpha = C_\beta^T T_{\beta\times\alpha}^L I_\alpha = -Z_{ji}^B B_i^H,$$

where we used that $T_{\beta\times\alpha}^L C_\alpha^T = C_\beta^T T_{\beta\times\alpha}^L$ for $\beta \geq \alpha$. Similarly for the (i, j) block,

$$A_i^H Z_{ij}^A = C_\alpha^T S_{\alpha\times\beta}^L = I_\alpha C_\alpha^T S_{\alpha\times\beta}^L = -B_i^H Z_{ij}^B$$

and

$$Z_{ij}^A A_j^H = S_{\alpha\times\beta}^L C_\beta^T = S_{\alpha\times\beta}^L C_\beta^T I_\beta = C_\alpha^T S_{\alpha\times\beta}^L I_\beta = -Z_{ij}^B B_j^H.$$

Here we used that $S_{\alpha\times\beta}^L C_\beta^T = C_\alpha^T S_{\alpha\times\beta}^L$ for $\beta \geq \alpha$.

Since the (i,i), (i,j), and (j,i) blocks of $Z_A - \lambda Z_B$ have α parameters each and the (j,j) block has β parameters, the total number of parameters in $Z_A - \lambda Z_B$ is equal to $\mathrm{cod}(J_\alpha(0) \oplus J_\beta(0)) = \beta + 3\alpha$.

$\mathbf{N}_\alpha \oplus \mathbf{N}_\beta$: Since there is a symmetry between $J_k(0) = C_k - \lambda I_k$ and $N_k = I_k - \lambda C_k$ and there is a corresponding symmetry between blocks in $Z_A - \lambda Z_B$ for $J_k(0)$ and N_k blocks, the proof for $N_\alpha \oplus N_\beta$ is similar to the case $J_\alpha(0) \oplus J_\beta(0)$.

$\mathbf{J}_\alpha(\gamma) \oplus \mathbf{J}_\beta(\gamma)$: Here the (j,i) block and the (i,j) block are defined similarly (see Table 5.1), and therefore it is sufficient to prove one of them with no constraints on α and β. We note that $J_k(\gamma) = \gamma I_k + C_k - \lambda I_k$. We show that the first and second conditions hold for $Z^A_{ji} = D_{\beta\times\alpha}[q]$ and $Z^B_{ji} = E_{\beta\times\alpha}[q]$ for $q = 1, \ldots, \min\{\alpha, \beta\}$. *First* condition:

$$A_j^H Z^A_{ji} = (\gamma I_\beta + C_\beta)^H D_{\beta\times\alpha}[q] = \overline{\gamma} D_{\beta\times\alpha}[q] + C_\beta^T D_{\beta\times\alpha}[q].$$

Remember that $D_{\beta\times\alpha}[q]$ has all elements zero, except for the q lower left diagonals, where all elements in each diagonal are identical and defined by the element in the first column. For $q = 1$ the proof is trivial. For $q > 1$, $A_j^H Z^A_{ji}$ gives the following matrix. All diagonals starting at position $(u, 1)$ for $1 \le u \le \beta - q$ are zero. The elements in the diagonal starting at position $(\beta - q + 1, 1)$ are $\overline{\gamma} d_q$, which by definition is equal to $-e_q^*$, which in turn defines the corresponding diagonal in $-E_{\beta\times\alpha}[q]$. The elements in the diagonals starting at positions $(\beta - u + 1, 1)$, where $1 \le u < q$, are equal to $\overline{\gamma} d_u + d_{u+1}$. Since d_{u+1} is defined as $-\overline{\gamma} d_u - e_u$, the elements in these diagonals are equal to $-e_u$, which defines the elements in the corresponding diagonals in $-E_{\beta\times\alpha}[q]$. Since $-E_{\beta\times\alpha}[q] = -B_j^H Z^B_{ji}$, we have proved the first condition.

Second condition: Since $D_{\beta\times\alpha}[q]$ only has $q \le \min\{s, t\}$ nonzero diagonals in the lower left corner of the matrix, a shift of rows downward gives the same result as a shift of columns leftward, i.e., $C_\beta^T D_{\beta\times\alpha}[q] = D_{\beta\times\alpha}[q] C_\alpha^T$. Using information from the first part, we obtain

$$\begin{aligned} Z^A_{ji} A_i^H &= D_{\beta\times\alpha}[q](\gamma I_\alpha + C_\alpha)^H = \overline{\gamma} D_{\beta\times\alpha}[q] + D_{\beta\times\alpha}[q] C_\alpha^T = \overline{\gamma} D_{\beta\times\alpha}[q] + C_\beta^T D_{\beta\times\alpha}[q] \\ &= A_j^H Z^A_{ji} = -E_{\beta\times\alpha}[q] = -Z^B_{ji} B_i^H \end{aligned}$$

since B_i is the identity matrix.

Also here, the number of parameters in $Z^A_{jj} - \lambda Z^B_{jj}$ is β, and there are α parameters in each of the other three blocks, giving $\beta + 3\alpha$ in total.

Even though the (i,i), (j,i), (i,j), and (j,j) blocks look rather complicated, they reduce for $\gamma = 0$ to the corresponding blocks for $J_\alpha(0) \oplus J_\beta(0)$ in Table 5.1.

$\mathbf{L}_\alpha \oplus \mathbf{L}_\beta$: Here we use $L_k = \hat{G}_k - \lambda G_k$. *First* condition for the (j,i) block:

$$A_j^H Z^A_{ji} = \hat{G}_\beta^T S^B_{\beta\times(\alpha+1)} = \begin{bmatrix} 0 \\ S^B_{\beta\times(\alpha+1)} \end{bmatrix} = \begin{bmatrix} C_\beta^T S^B_{\beta\times(\alpha+1)} \\ 0 \end{bmatrix} = G_\beta^T C_\beta^T S^B_{\beta\times(\alpha+1)} = -B_j^H Z^B_{ji}.$$

Second condition for the (j,i) block:

$$Z^A_{ji} A_i^H = S^B_{\beta\times(\alpha+1)} \hat{G}_\beta^T = \begin{bmatrix} 0 \\ S^B_{\beta\times(\alpha+1)} \end{bmatrix} = \begin{bmatrix} C_\beta^T S^B_{\beta\times(\alpha+1)} \\ 0 \end{bmatrix} = C_\beta^T S^B_{\beta\times(\alpha+1)} G_\beta^T = -Z^B_{ji} B_i^H.$$

Since the contribution from $L_\alpha \oplus L_\beta$ to the codimension is $\beta - \alpha - 1$ and the (j,i) block has $\beta - \alpha - 1$ independent parameters, we deduce that all other blocks in $Z_A - \lambda Z_B$ are zero.

$\mathbf{L}_\alpha^\mathbf{T} \oplus \mathbf{L}_\beta^\mathbf{T}$: Since this case is just the transpose of $L_\alpha \oplus L_\beta$, the proof is almost the same, and therefore we omit the technical details here.

So far we have proved all cases where both blocks are of the same type. Since the diagonal blocks in $Z_A - \lambda Z_B$ always correspond to such cases (see Table 5.3 for the number of parameters in these blocks), from now on we only have to consider the (i,j) and (j,i) blocks, where $i \neq j$ for the remaining cases.

$\mathbf{L}_\alpha \oplus \mathbf{J}_\beta(\mathbf{0})$: *First* condition for the (j,i) block:

$$A_j^H Z_{ji}^A = C_\beta^T S_{\beta\times(\alpha+1)}^L = I_\beta C_\beta^T S_{\beta\times(\alpha+1)}^L = -B_j^H Z_{ji}^B.$$

Second condition for the (j,i) block:

$$Z_{ji}^A A_i^H = S_{\beta\times(\alpha+1)}^L \hat{G}_\alpha^T = C_\beta^T S_{\beta\times(\alpha+1)}^L G_\alpha^T = -Z_{ji}^B B_i^H.$$

The (i,i) and (j,j) blocks contribute with zero and β parameters, respectively. Since the (j,i) block gives another β parameters, we have found all 2β parameters, and therefore it follows that $Z_{ij}^A = \lambda Z_{ij}^B = 0$.

$\mathbf{L}_\alpha \oplus \mathbf{J}_\beta(\gamma)$: *First* condition for the (j,i) block:

$$A_j^H Z_{ji}^A = (\gamma I_\beta + C_\beta)^H F_{\beta\times(\beta+2)}^D G_{\alpha+1}^T = \overline{\gamma} F_{\beta\times(\beta+2)}^D G_{\alpha+1}^T + C_\beta^T F_{\beta\times(\beta+2)}^D G_{\alpha+1}^T.$$

By inspection we see that the (u,v) element of this matrix is $\overline{\gamma} f_{u,v}^d + f_{u-1,v}^d$ if $u > 1$ and $\overline{\gamma} f_{u,v}^d$ if $u = 1$ (where $f_{u,v}^d$ denotes the (u,v) element of F^D). The right-hand side of the same condition is

$$-B_j^H Z_{ji}^B = I_\beta F_{\beta\times(\beta+2)}^D \hat{G}_{\alpha+1}^T,$$

which simply is the β leftmost columns of $F_{\beta\times(\beta+2)}^D$. The (u,v) element of this matrix is then $f_{u,v+1}^d$, which is defined as $\overline{\gamma} f_{u,v}^d + f_{u-1,v}^d$ if $u > 1$ and $\overline{\gamma} f_{u,v}^d$ if $u = 1$.

Second condition for the (j,i) block:

$$Z_{ji}^A A_i^H \hat{G}_\alpha^T = F_{\beta\times(\alpha+2)}^D G_{\alpha+1}^T \hat{G}_\alpha^T = F_{\beta\times(\alpha+2)}^D \begin{bmatrix} 0 \\ I_\alpha \\ 0 \end{bmatrix} = F_{\beta\times(\alpha+2)}^D \hat{G}_{\alpha+1} G_\alpha^T = -Z_{ji}^B B_i^H.$$

As in the previous case, the (i,i) and (j,j) blocks contribute with zero and β parameters, respectively. Since the (j,i) block gives the remaining β parameters, the (i,j) block is the zero pencil.

Notably, for $\gamma = 0$, the "monstrous" (j,i) block reduces to the (j,i) block for $L_\alpha \oplus J_\beta(0)$ in Table 5.1.

$\mathbf{L}_\alpha \oplus \mathbf{N}_\beta$: *First* condition for the (j,i) block:

$$A_j^H Z_{ji}^A = I_\beta C_\beta^T H_{\beta\times(\alpha+1)}^L = C_\beta^T H_{\beta\times(\alpha+1)}^L = -B_j^H Z_{ji}^B.$$

Second condition for the (j,i) block:

$$Z_{ji}^A A_i^H \hat{G}_\alpha^T = C_\beta^T H_{\beta\times(\alpha+1)}^L = \begin{bmatrix} 0 \\ H_{(\beta-1)\times\alpha}^L \end{bmatrix} = H_{\beta\times(\alpha+1)}^L G_\alpha^T = -Z_{ji}^B B_i^H.$$

Also here, the (i,i) and (j,j) blocks contribute with zero and β parameters, respectively. Since the (j,i) block gives the remaining β parameters, the (i,j) block is the zero pencil.

$\mathbf{L}_\alpha \oplus \mathbf{L}_\beta^{\mathbf{T}}$: For this case the (i,i) and (j,j) blocks are zero pencils. *First* condition for the (j,i) block:

$$\begin{aligned} A_j^H Z_{ji}^A &= \hat{G}_\beta G_{\beta+1} H_{(\beta+2)\times(\alpha+1)} = [0\ I_\beta\ 0] H_{(\beta+2)\times(\alpha+1)} \\ &= G_\beta \hat{G}_{\beta+1} H_{(\beta+2)\times(\alpha+1)} = -B_j^H Z_{ji}^B. \end{aligned}$$

Second condition for the (j,i) block:

$$Z_{ji}^A A_i^H = G_{\beta+1} H_{(\beta+2)\times(\alpha+1)} \hat{G}_\alpha^T,$$

which is a matrix consisting of the $\beta+1$ first rows and α last columns of $H_{(\beta+2)\times(\alpha+1)}$. This matrix is identical to the one given by the $\beta+1$ last rows and α first columns of $H_{(\beta+2)\times(\alpha+1)}$, i.e.,

$$\hat{G}_{\beta+1} H_{(\beta+2)\times(\alpha+1)} G_\alpha^T = -Z_{ji}^B B_i^H.$$

Since this block has all $\alpha+\beta+2$ parameters, it follows that the (i,j) block is the zero pencil.

$\mathbf{J}_\alpha(\mathbf{0}) \oplus \mathbf{L}_\beta^{\mathbf{T}}$: *First* condition for the (j,i) block:

$$A_j^H Z_{ji}^A = \hat{G}_\beta H_{(\beta+1)\times\alpha}^U,$$

which simply is the last β rows in $H_{(\beta+1)\times\alpha}^U$. Another way to construct this matrix is to shift the columns in $H_{(\beta+1)\times\alpha}^U$ one column leftward and pick the β first columns of the matrix, which can be written as

$$G_\beta H_{(\beta+1)\times\alpha}^U C_\alpha^T = -B_j^H Z_{ji}^B.$$

Second condition for the (j,i) block:

$$Z_{ji}^A A_i^H = H_{(\beta+1)\times\alpha}^U C_\alpha^T = H_{(\beta+1)\times\alpha}^U C_\alpha^T I_\alpha = -Z_{ji}^B B_i^H.$$

The (i,i) and (j,j) blocks contribute with α and zero parameters, respectively. Since the (j,i) block gives another α parameters, we conclude that the (i,j) block is the zero pencil.

$\mathbf{J}_\alpha(\gamma) \oplus \mathbf{L}_\beta^{\mathbf{T}}$: Since the proof for this case is similar to the one for the case $L_\alpha \oplus J_\beta(\gamma)$, we omit the technical details here. It follows that for $\gamma = 0$, the (j,i) block reduces to the (j,i) block for $J_\alpha(0) \oplus L_\beta^T$ in Table 5.1.

$\mathbf{N}_\alpha \oplus \mathbf{L}_\beta^{\mathbf{T}}$: *First* condition for the (j,i) block:

$$A_j^H Z_{ji}^A = \hat{G}_\beta T_{(\beta+1)\times\alpha}^L C_\alpha^T,$$

which is the last β rows in $T_{(\beta+1)\times\alpha}^L$ shifted one column leftward. This matrix is identical to the one given by the β first rows in $T_{(\beta+1)\times\alpha}^L$, which is

$$G_\beta T_{(\beta+1)\times\alpha}^L = -B_j^H Z_{ji}^B.$$

Second condition for the (j,i) block:

$$Z_{ji}^A A_i^H = T_{(\beta+1)\times\alpha}^L C_\alpha^T I_\alpha = T_{(\beta+1)\times\alpha}^L C_\alpha^T = -Z_{ji}^B B_i^H.$$

The (i,i) and (j,j) blocks in $Z_A - \lambda Z_B$ contribute with α and zero parameters, respectively. Since the (j,i) block gives another α parameters, we conclude that the (i,j) block is the zero pencil.

$\mathbf{J}_\alpha(\mathbf{0}) \oplus \mathbf{J}_\beta(\gamma)$, $\mathbf{J}_\alpha(\mathbf{0}) \oplus \mathbf{N}_\beta$, $\mathbf{J}_\alpha(\gamma_1) \oplus \mathbf{J}_\beta(\gamma_2)$, and $\mathbf{J}_\alpha(\gamma) \oplus \mathbf{N}_\beta$: In these four cases the (i,i) and (j,j) blocks contribute with α and β parameters, respectively, and therefore the (j,i) and (i,j) blocks are zero pencils.

Since we have considered all possible cases of M_i and M_j blocks, the proof is complete. □

6. Applications and examples.

6.1. Some examples of versal deformations of matrix pencils in KCF. In the following we show three examples of versal deformations of matrix pencils. For the 7×8 pencil $A - \lambda B = L_2 \oplus J_2(0) \oplus J_3(0)$ with codimension 14, the 14-parameter versal deformation $\mathcal{V}(p) = A - \lambda B + Z_A - \lambda Z_B$, where $Z_A - \lambda Z_B \in \mathrm{nor}(A - \lambda B)$, is given by

$$Z_A = \left[\begin{array}{ccc|cc|ccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline p_1 & 0 & 0 & p_6 & 0 & p_{10} & 0 & 0 \\ p_2 & p_1 & 0 & p_7 & p_6 & p_{11} & p_{10} & 0 \\ \hline p_3 & 0 & 0 & 0 & 0 & p_{12} & 0 & 0 \\ p_4 & p_3 & 0 & p_8 & 0 & p_{13} & p_{12} & 0 \\ p_5 & p_4 & p_3 & p_9 & p_8 & p_{14} & p_{13} & p_{12} \end{array}\right]$$

and

$$Z_B = \left[\begin{array}{ccc|cc|ccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -p_1 & 0 & 0 & -p_6 & 0 & -p_{10} & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -p_3 & 0 & 0 & 0 & 0 & -p_{12} & 0 & 0 \\ -p_4 & -p_3 & 0 & -p_8 & 0 & -p_{13} & -p_{12} & 0 \end{array}\right].$$

For the 3×4 pencil $A - \lambda B = L_1 \oplus J_2(\gamma)$ with codimension 4, the four-parameter versal deformation $\mathcal{V}(p) = A - \lambda B + Z_A - \lambda Z_B$, where $Z_A - \lambda Z_B \in \mathrm{nor}(A - \lambda B)$, is given by

$$Z_A = \left[\begin{array}{cc|cc} 0 & 0 & 0 & 0 \\ \hline p_1 & p_1\overline{\gamma} & p_3(|\gamma|^2+1) & 0 \\ p_2 - p_1\frac{2\gamma}{|\gamma|^2+1} & p_2\overline{\gamma} - p_1\frac{|\gamma|^2-1}{|\gamma|^2+1} & -p_3\gamma + p_4 & p_3(|\gamma|^2+1) \end{array}\right]$$

and

$$Z_B = \left[\begin{array}{cc|cc} 0 & 0 & 0 & 0 \\ \hline -p_1\overline{\gamma} & -p_1\overline{\gamma}^2 & -p_3(|\gamma|^2\overline{\gamma}+\overline{\gamma}) & 0 \\ -p_2\overline{\gamma} + p_1\frac{|\gamma|^2-1}{|\gamma|^2+1} & -p_2\overline{\gamma}^2 - p_1\frac{2\overline{\gamma}}{|\gamma|^2+1} & -p_3 - p_4\overline{\gamma} & -p_3(|\gamma|^2\overline{\gamma}+\overline{\gamma}) \end{array}\right].$$

For the 11×11 pencil $A - \lambda B = L_1 \oplus J_3(0) \oplus N_4 \oplus L_2^T$ with codimension 26, the 26-parameter versal deformation $\mathcal{V}(p) = A - \lambda B + Z_A - \lambda Z_B$, where $Z_A - \lambda Z_B \in$

$\mathrm{nor}(A-\lambda B)$, is given by

$$Z_A = \left[\begin{array}{cc|ccc|ccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline p_1 & 0 & p_{13} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_2 & p_1 & p_{14} & p_{13} & 0 & 0 & 0 & 0 & 0 & 0 \\ p_3 & p_2 & p_{15} & p_{14} & p_{13} & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & p_4 & 0 & 0 & 0 & p_{19} & 0 & 0 & 0 & 0 \\ p_4 & p_5 & 0 & 0 & 0 & p_{20} & p_{19} & 0 & 0 & 0 \\ p_5 & p_6 & 0 & 0 & 0 & p_{21} & p_{20} & p_{19} & 0 & 0 \\ \hline p_8 & p_9 & p_{18} & p_{17} & p_{16} & p_{23} & 0 & 0 & 0 & 0 \\ p_9 & p_{10} & p_{17} & p_{16} & 0 & p_{24} & p_{23} & 0 & 0 & 0 \\ p_{10} & p_{11} & p_{16} & 0 & 0 & p_{25} & p_{24} & p_{23} & 0 & 0 \end{array}\right]$$

and

$$Z_B = \left[\begin{array}{cc|ccc|ccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -p_1 & 0 & -p_{13} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -p_2 & -p_1 & -p_{14} & -p_{13} & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & -p_4 & 0 & 0 & 0 & -p_{19} & 0 & 0 & 0 & 0 \\ -p_4 & -p_5 & 0 & 0 & 0 & -p_{20} & -p_{19} & 0 & 0 & 0 \\ -p_5 & -p_6 & 0 & 0 & 0 & -p_{21} & -p_{20} & -p_{19} & 0 & 0 \\ -p_6 & -p_7 & 0 & 0 & 0 & -p_{22} & -p_{21} & -p_{20} & -p_{19} & 0 & 0 \\ \hline -p_9 & -p_{10} & -p_{17} & -p_{16} & 0 & -p_{24} & -p_{23} & 0 & 0 & 0 & 0 \\ -p_{10} & -p_{11} & -p_{16} & 0 & 0 & -p_{25} & -p_{24} & -p_{23} & 0 & 0 & 0 \\ -p_{11} & -p_{12} & 0 & 0 & 0 & -p_{26} & -p_{25} & -p_{24} & -p_{23} & 0 & 0 \end{array}\right].$$

6.2. Versal deformations of the set of 2×3 matrix pencils. In [15], the algebraic and geometric characteristics of the set of 2×3 matrix pencils were examined in full detail, including the complete closure hierarchy. There, all nonzero and finite eigenvalues were considered as unspecified. R_2 was used to denote a 2×2 block with nonzero finite eigenvalues, i.e., any of the three structures $J_1(\alpha)\oplus J_1(\beta)$, $J_1(\alpha)\oplus J_1(\alpha)$, and $J_2(\alpha)$, where $\alpha, \beta \neq \{0, \infty\}$. However, in the context of versal deformations all these forms are considered separately and with the eigenvalues specified (known). Consequently, we now have 20 different Kronecker structures to investigate. For example, the versal deformation of $A - \lambda B = L_0 \oplus J_2(\gamma)$, $\gamma \neq \{0, \infty\}$, is found by computing $Z_A - \lambda Z_B =$

(6.1)
$$\left[\begin{array}{ccc} p_1 + \lambda\overline{\gamma}\, p_1 & p_3(|\gamma|^2+1) + p_3(|\gamma|^2\overline{\gamma}+\overline{\gamma}) & 0 \\ p_2 - \frac{p_1\gamma}{|\gamma|^2+1} + \lambda(p_2\overline{\gamma} + \frac{p_1}{|\gamma|^2+1}) & -p_3\gamma + p_4 + \lambda(p_3 + p_4\overline{\gamma}) & p_3(|\gamma|^2+1) + p_3(|\gamma|^2\overline{\gamma}+\overline{\gamma}) \end{array}\right].$$

In Table 6.1 we show the versal deformations for all different Kronecker structures for this set of matrix pencils. The different structures are displayed in increasing codimension order.

6.2.1. Using GUPTRI in a random walk in tangent and normal directions of nongeneric pencils. To illustrate how perturbations in the tangent space and the normal space affect the Kronecker structure computed by a staircase algorithm, we have performed a set of tests on nongeneric 2×3 matrix pencils. Since the

TABLE 6.1
Versal deformations $\mathcal{V}(p) = A - \lambda B + Z_A - \lambda Z_B$ of the set of 2×3 matrix pencils.

KCF	$A - \lambda B$	$Z_A - \lambda Z_B$
L_2	$\begin{bmatrix} -\lambda & 1 & 0 \\ 0 & -\lambda & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$
$L_1 \oplus J_1(\gamma)$	$\begin{bmatrix} -\lambda & 1 & 0 \\ 0 & 0 & \gamma - \lambda \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ p_1 + \lambda\overline{\gamma}\, p_1 & \overline{\gamma}\, p_1 + \lambda\overline{\gamma}^2 p_1 & p_2 + \lambda\overline{\gamma}\, p_2 \end{bmatrix}$
$L_1 \oplus J_1(0)$	$\begin{bmatrix} -\lambda & 1 & 0 \\ 0 & 0 & -\lambda \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ p_1 & 0 & p_2 \end{bmatrix}$
$L_1 \oplus N_1$	$\begin{bmatrix} -\lambda & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & \lambda p_1 & \lambda p_2 \end{bmatrix}$
$L_0 \oplus J_1(\gamma_1) \oplus J_1(\gamma_2)$	$\begin{bmatrix} 0 & \gamma_1 - \lambda & 0 \\ 0 & 0 & \gamma_2 - \lambda \end{bmatrix}$	$\begin{bmatrix} p_1 + \lambda\overline{\gamma}_1 p_1 & p_3 + \lambda\overline{\gamma}_1 p_3 & 0 \\ p_2 + \lambda\overline{\gamma}_2 p_2 & 0 & p_4 + \lambda\overline{\gamma}_2 p_4 \end{bmatrix}$
$L_0 \oplus J_2(\gamma)$	$\begin{bmatrix} 0 & \gamma - \lambda & 1 \\ 0 & 0 & \gamma - \lambda \end{bmatrix}$	See (6.1)
$L_0 \oplus 2J_1(\gamma)$	$\begin{bmatrix} 0 & \gamma - \lambda & 0 \\ 0 & 0 & \gamma - \lambda \end{bmatrix}$	$\begin{bmatrix} p_1 + \lambda\overline{\gamma}\, p_1 & p_3 + \lambda\overline{\gamma}\, p_3 & p_5 + \lambda\overline{\gamma}\, p_5 \\ p_2 + \lambda\overline{\gamma}\, p_2 & p_4 + \lambda\overline{\gamma}\, p_4 & p_6 + \lambda\overline{\gamma}\, p_6 \end{bmatrix}$
$L_0 \oplus J_1(0) \oplus J_1(\gamma)$	$\begin{bmatrix} 0 & -\lambda & 0 \\ 0 & 0 & \gamma - \lambda \end{bmatrix}$	$\begin{bmatrix} p_1 & p_3 & 0 \\ p_2 + \lambda\overline{\gamma}\, p_2 & 0 & p_4 + \lambda\overline{\gamma}\, p_4 \end{bmatrix}$
$L_0 \oplus J_1(\gamma) \oplus N_1$	$\begin{bmatrix} 0 & \gamma - \lambda & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} p_1 + \lambda\overline{\gamma}\, p_1 & p_3 + \lambda\overline{\gamma}\, p_3 & 0 \\ \lambda p_2 & 0 & \lambda p_4 \end{bmatrix}$
$L_0 \oplus J_2(0)$	$\begin{bmatrix} 0 & -\lambda & 1 \\ 0 & 0 & -\lambda \end{bmatrix}$	$\begin{bmatrix} p_1 & p_3 & 0 \\ p_2 + \lambda p_1 & p_4 + \lambda p_3 & p_3 \end{bmatrix}$
$L_0 \oplus N_2$	$\begin{bmatrix} 0 & 1 & -\lambda \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} \lambda p_1 & \lambda p_3 & 0 \\ p_1 + \lambda p_2 & p_3 + \lambda p_4 & \lambda p_3 \end{bmatrix}$
$L_0 \oplus J_1(0) \oplus N_1$	$\begin{bmatrix} 0 & -\lambda & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} p_1 & p_3 & 0 \\ \lambda p_2 & 0 & \lambda p_4 \end{bmatrix}$
$L_0 \oplus L_1 \oplus L_0^T$	$\begin{bmatrix} 0 & -\lambda & 1 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ p_1 + \lambda p_2 & p_3 + \lambda p_4 & p_4 + \lambda p_5 \end{bmatrix}$
$L_0 \oplus 2J_1(0)$	$\begin{bmatrix} 0 & -\lambda & 0 \\ 0 & 0 & -\lambda \end{bmatrix}$	$\begin{bmatrix} p_1 & p_3 & p_5 \\ p_2 & p_4 & p_6 \end{bmatrix}$
$L_0 \oplus 2N_1$	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} \lambda p_1 & \lambda p_3 & \lambda p_5 \\ \lambda p_2 & \lambda p_4 & \lambda p_6 \end{bmatrix}$
$2L_0 \oplus L_1^T$	$\begin{bmatrix} 0 & 0 & -\lambda \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} p_1 + \lambda p_2 & p_4 + \lambda p_5 & 0 \\ p_2 + \lambda p_3 & p_5 + \lambda p_6 & 0 \end{bmatrix}$
$2L_0 \oplus J_1(\gamma) \oplus L_0^T$	$\begin{bmatrix} 0 & 0 & \gamma - \lambda \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} p_1 + \lambda\overline{\gamma}\, p_1 & p_4 + \lambda\overline{\gamma}\, p_4 & p_7 + \lambda\overline{\gamma}\, p_7 \\ p_2 + \lambda p_3 & p_5 + \lambda p_6 & p_8 + \lambda\overline{\gamma}\, p_8 \end{bmatrix}$
$2L_0 \oplus J_1(0) \oplus L_0^T$	$\begin{bmatrix} 0 & 0 & -\lambda \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} p_1 & p_4 & p_7 \\ p_2 + \lambda p_3 & p_5 + \lambda p_6 & p_8 \end{bmatrix}$
$2L_0 \oplus N_1 \oplus L_0^T$	$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} \lambda p_1 & \lambda p_4 & \lambda p_7 \\ p_2 + \lambda p_3 & p_5 + \lambda p_6 & \lambda p_8 \end{bmatrix}$
$3L_0 \oplus 2L_0^T$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} p_1 + \lambda p_2 & p_5 + \lambda p_6 & p_9 + \lambda p_{10} \\ p_3 + \lambda p_4 & p_7 + \lambda p_8 & p_{11} + \lambda p_{12} \end{bmatrix}$

staircase algorithm considers all nonzero finite eigenvalues as unspecified, we have not included these cases in the test.

For the remaining 12 nongeneric cases a random perturbation $E_A - \lambda E_B$, with entries e_{ij}^a, e_{ij}^b, has been decomposed into two parts $T_A - \lambda T_B \in \tan(A - \lambda B)$ and

$Z_A - \lambda Z_B \in \mathrm{nor}(A - \lambda B)$ such that

$$E_A = T_A + Z_A \quad \text{and} \quad E_B = T_B + Z_B.$$

We illustrate the decomposition of $E_A - \lambda E_B$ with $A - \lambda B = L_0 \oplus J_2(0)$. From Table 6.1 we get

$$Z_A = \begin{bmatrix} p_1 & p_3 & 0 \\ p_2 & p_4 & p_3 \end{bmatrix}, \quad Z_B = \begin{bmatrix} 0 & 0 & 0 \\ -p_1 & -p_3 & 0 \end{bmatrix}.$$

Let $T_A - \lambda T_B = (E_A - \lambda E_B) - (Z_A - \lambda Z_B)$. Now, the parameters p_i are determined by computing the component of $E_A - \lambda E_B$ in each of the four orthogonal (but not orthonormal) directions that span the normal space:

$$\begin{aligned}
Z_1 &= \frac{1}{2}\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}\right), \\
Z_2 &= 1\left(\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right), \\
Z_3 &= \frac{1}{3}\left(\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}\right), \\
Z_4 &= 1\left(\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right).
\end{aligned}$$

We conclude that

$$p_1 = \frac{e_{11}^a - e_{21}^b}{2}, \quad p_2 = e_{21}^a, \quad p_3 = \frac{e_{12}^a + e_{23}^a - e_{22}^b}{3}, \quad p_4 = e_{22}^a.$$

It is easily verified that $\langle T_A - \lambda T_B, Z_A - \lambda Z_B \rangle = 0$.

GUPTRI [11, 12] has been used to compute the Kronecker structure of the perturbed pencils $A - \lambda B + \epsilon(E_A - \lambda E_B)$, $A - \lambda B + \epsilon(Z_A - \lambda Z_B)$, and $A - \lambda B + \epsilon(T_A - \lambda T_B)$ for $\epsilon = 10^{-16}, 10^{-15}, \ldots, 10^{0}$. We investigate how far we can move in the tangent and normal directions before GUPTRI reports the generic Kronecker structure.

The procedure has been repeated for all cases and for 100 random perturbations (E_A, E_B), where $\|(E_A, E_B)\|_F = 1$ and $\|E_A\|_F = \|E_B\|_F$. The entries of (E_A, E_B) are uniformly distributed in $(-0.5, 0.5)$. For each case and for each perturbation $E_A - \lambda E_B$ we record the size of ϵ when GUPTRI reports the generic Kronecker structure. In Table 6.2 we display the smallest, median, and maximum values of ϵ for the 100 random perturbations.

Entries marked + in Table 6.2 mean that the generic structure was not found for any size of the perturbations. All these results were for perturbations in $\tan(A - \lambda B)$, and they indicate that for these Kronecker structures there is little or no curvature in the orbit at this point (pencil) in this direction. Here the tangent directions are very close to $\mathrm{orbit}(A - \lambda B)$.

Notably, the results for the perturbations $\epsilon(E_A - \lambda E_B)$ are, except for one case, similar to the results for $\epsilon(Z_A - \lambda Z_B)$. This is natural since the perturbation $E_A - \lambda E_B$ implies a translation both in the tangent space and the normal space directions. The structure changes appear more rapidly in the normal space, i.e., for smaller ϵ. Our computational results extend the cone example in section 1.3 to 2×3 matrix pencils.

Why is the smallest perturbation $10^{-16}(Z_A - \lambda Z_B)$ sufficient to find the generic structure for the three cases $L_0 \oplus 2J_1(0)$, $L_0 \oplus 2N_1$, and $3L_0 \oplus 2L_0^T$? The explanation is connected to the procedure for determining the numerical rank of matrices.

TABLE 6.2
How far we can move in tangent and normal directions before nongeneric 2×3 matrix pencils turn generic.

		$\epsilon(Z_A - \lambda Z_B)$			$\epsilon(T_A - \lambda T_B)$		
$A - \lambda B$	$\mathrm{cod}(A - \lambda B)$	$\epsilon_{\min}$	$\epsilon_{\mathrm{median}}$	$\epsilon_{\max}$	$\epsilon_{\min}$	$\epsilon_{\mathrm{median}}$	$\epsilon_{\max}$
$L_1 \oplus J_1(0)$	2	10^{-4}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{-1}
$L_1 \oplus N_1$	2	10^{-4}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
$L_0 \oplus J_2(0)$	4	10^{-4}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
$L_0 \oplus N_2$	4	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{-1}
$L_0 \oplus J_1(0) \oplus N_1$	4	10^{-4}	10^{-4}	10^{-2}	10^{-2}	10^{-1}	10^{0}
$L_0 \oplus L_1 \oplus L_0^T$	5	10^{-4}	10^{-4}	10^{-2}	10^{-2}	10^{-1}	10^{0}
$L_0 \oplus 2J_1(0)$	6	10^{-16}	10^{-16}	10^{-16}	+	+	+
$L_0 \oplus 2N_1$	6	10^{-16}	10^{-16}	10^{-16}	+	+	+
$2L_0 \oplus L_1^T$	6	10^{-4}	10^{-4}	10^{-2}	+	+	+
$2L_0 \oplus J_1(0) \oplus L_0^T$	8	10^{-5}	10^{-4}	10^{-1}	+	+	+
$2L_0 \oplus N_1 \oplus L_0^T$	8	10^{-4}	10^{-4}	10^{-3}	+	+	+
$3L_0 \oplus 2L_0^T$	12	10^{-16}	10^{-16}	10^{-16}	+	+	+

`GUPTRI` has two input parameters, `EPSU` and `GAP`, which are used to make rank decisions to determine the Kronecker structure of an input pencil $A - \lambda B$. Inside `GUPTRI` the absolute tolerances $\texttt{EPSUA} = \|A\|_E \cdot \texttt{EPSU}$ and $\texttt{EPSUB} = \|B\|_E \cdot \texttt{EPSU}$ are used in all rank decisions, where the matrices A and B, respectively, are involved. Suppose the singular values of A are computed in increasing order, i.e., $0 \le \sigma_1 \le \sigma_2 \le \cdots \le \sigma_k \le \sigma_{k+1} \le \cdots$; then all singular values $\sigma_k <$ `EPSUA` are interpreted as zeros. The rank decision is made more robust in practice: if $\sigma_k <$ `EPSUA` but $\sigma_{k+1} \ge$ `EPSUA`, `GUPTRI` insists on a gap between the two singular values such that $\sigma_{k+1}/\sigma_k \ge$ `GAP`. If $\sigma_{k+1}/\sigma_k <$ `GAP`, σ_{k+1} is also treated as zero. This process is repeated until an appreciable gap between the zero and nonzero singular values is obtained. In all of our tests we have used $\texttt{EPSU} = 10^{-8}$ and $\texttt{GAP} = 1000.0$.

For the most nongeneric case $3L_0 \oplus 2L_0^T$, both the A-part and the B-part are zero matrices giving $\texttt{EPSUA} = \texttt{EPSUB} = 0$, which in turn leads to the decision that a full rank perturbation $E_A - \lambda E_B$ times a very small ϵ is interpreted as a generic pencil. For the other two cases, either the A-part or the B-part is full rank and the other part is a zero matrix, which accordingly is interpreted to have full rank already for the smallest perturbation.

6.2.2. Versal deformations and minimal perturbations for changing a nongeneric structure. In the following we illustrate how versal deformations are useful in the understanding of the relations between the different structures by looking at requirements on perturbations to (A, B) for changing the Kronecker structure. Assume that we have the following matrix pencil with the Kronecker structure $L_1 \oplus J_1(0)$:

$$A - \lambda B = \begin{bmatrix} -\epsilon_1\lambda & \epsilon_2 & 0 \\ 0 & 0 & -\epsilon_3\lambda \end{bmatrix} \quad \text{and} \quad Z_A - \lambda Z_B = \begin{bmatrix} 0 & 0 & 0 \\ p_1 & 0 & p_2 \end{bmatrix}. \tag{6.2}$$

It was shown in [15] that $L_1 \oplus J_1(0)$ with codimension 2 is in the closure of orbit$(L_1 \oplus J_1(\gamma))$ ($\gamma \neq \{0, \infty\}$ but otherwise unspecified) with codimension 1, which in turn is in the closure of orbit(L_2) (the generic KCF) with codimension 0. Notice that in Table 6.1, since γ is assumed specified, $L_1 \oplus J_1(\gamma)$ has two parameters (and codimension = 2). In the discussion that follows we assume that γ is finite and nonzero but unspecified.

We will now, for this example, illustrate how perturbations in the normal space directions can be used to find more generic Kronecker structures (going upward in the Kronecker structure hierarchy) and how we can perturb the elements in $A - \lambda B$ to find less generic matrix pencils. Since the space spanned by $Z_A - \lambda Z_B$ is the normal space, we must always first hit a more generic pencil when we move infinitesimally in normal space directions.

The KCF remains unchanged as long as $p_1 = p_2 = 0$, but for $p_1 = 0$ and $p_2 \neq 0$, the KCF is changed into $L_1 \oplus J_1(\gamma)$ (with $\gamma = p_2$). That is, by adding a component in a normal space direction, we find a more generic pencil in the closure hierarchy. Notably, the size of the required perturbation is equal to the smallest size of an eigenvalue to be interpreted as nonzero. By choosing p_1 nonzero (and p_2 arbitrary), the resulting pencil will be generic with the KCF L_2.

To find a less generic structure, we may proceed in one of the following ways.

1. Find a less generic structure in the closure of orbit$(L_1 \oplus J_1(0))$.
2. Go upward in the closure hierarchy to a more generic structure and then look in that orbit's closure for a less generic structure.

We know from the investigation in [15] that all structures with higher codimension than $A - \lambda B = L_1 \oplus J_1(0)$ include an L_0 block in their Kronecker structures, which in turn implies that A and B must have a common column nullspace of at least dimension 1. Therefore, the smallest perturbation that turns $L_1 \oplus J_1(0)$ less generic is the smallest perturbation that reduces the rank of

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} 0 & \epsilon_2 & 0 \\ 0 & 0 & 0 \\ \epsilon_1 & 0 & 0 \\ 0 & 0 & \epsilon_3 \end{bmatrix}.$$

The size of the smallest rank-reducing perturbation is equal to the smallest of the singular values ϵ_1, ϵ_2, and ϵ_3. By just deleting one ϵ_i, the corresponding perturbed pencil is a less generic pencil within the closure of orbit$(L_1 \oplus J_1(0))$. These three cases correspond to approach 1 above. We summarize these perturbations and the perturbations in the normal space in Table 6.3. Notice that approach 2 will always require a perturbation larger than $\min\{\epsilon_i\}$.

Which of the nongeneric structures displayed in Table 6.3 is obtained by the smallest perturbation to $L_1 \oplus J_1(0)$? Mathematically, it is easy to see that the perturbations in the normal space always can be made smaller than a rank-reducing perturbation ϵ_i, since p_1 and p_2 are parameters that can be chosen arbitrarily small, e.g., smaller than $\min\{\epsilon_i\}$.

However, in finite-precision arithmetic, it is not clear that the smallest perturbation required to find another structure is in the normal direction. This can be illustrated by using `GUPTRI` to compute the Kronecker structures for $A - \lambda B$ as in (6.2) and perturbed as in Table 6.3. For `EPSU` $= 10^{-8}$, $\epsilon_2 = 1$, and $\epsilon_1 = \epsilon_3 = 10^{-10}$, `GUPTRI` uses different tolerances `EPSUA` $= 10^{-8}$ and `EPSUB` $= 10^{-18}$ for making rank decisions in A and B, respectively. It follows that for p_1 and p_2 of order 10^{-6}, `GUPTRI` still computes the Kronecker structure $L_1 \oplus J_1(0)$. However, if $p_1 = p_2 = 0$ and the B-part of the pencil is perturbed by ϵ_1 or ϵ_3, `GUPTRI` computes the less generic structures, just as shown in Table 6.3.

7. Conclusions. In this paper, we have obtained not only versal deformations for deformations of KCFs, but more importantly for our purposes, metrical information for the perturbation theory of matrix pencils relevant to the KCF. We demon-

TABLE 6.3
Perturbing $A - \lambda B$ (defined in (6.2)) yields the pencil $\tilde{A} - \lambda\tilde{B}$ with more or less generic structures. The codimension of the original orbit is 2.

$\|(\Delta A, \Delta B)\|_F$	$\tilde{A} - \lambda\tilde{B}$	KCF	$\operatorname{cod}(\tilde{A} - \lambda\tilde{B})$
p_1	$\begin{bmatrix} -\epsilon_1\lambda & \epsilon_2 & 0 \\ p_1 & 0 & -\epsilon_3\lambda \end{bmatrix}$	L_2	0
p_2	$\begin{bmatrix} -\epsilon_1\lambda & \epsilon_2 & 0 \\ 0 & 0 & p_2 - \epsilon_3\lambda \end{bmatrix}$	$L_1 \oplus J_1(p_2)$	1 (2)
ϵ_1	$\begin{bmatrix} 0 & \epsilon_2 & 0 \\ 0 & 0 & -\epsilon_3\lambda \end{bmatrix}$	$L_0 \oplus J_1(0) \oplus N_1$	4
ϵ_3	$\begin{bmatrix} -\epsilon_1\lambda & \epsilon_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$L_0 \oplus L_1 \oplus L_0^T$	5
ϵ_2	$\begin{bmatrix} -\epsilon_1\lambda & 0 & 0 \\ 0 & 0 & -\epsilon_3\lambda \end{bmatrix}$	$L_0 \oplus 2J_1(0)$	6

strated with numerical experiments in section 6 how this theory may be used in practice to see how computations are influenced by the geometry. In Part II of this paper, we will explore the stratification theory of matrix pencils with the goal of making algorithmic use of the lattice of orbits under the closure relationship [14].

Acknowledgments. We would like to thank Jim Demmel for conveying the message that the geometry of matrix and matrix pencil spaces influence perturbation theory and numerical algorithms. We further thank him, Richard Stanley, and David Vogan for many helpful discussions. In addition, the first author would like to thank the second two authors for their kind hospitality and support at Umeå University where this work and fruitful collaboration began.

REFERENCES

[1] V. I. ARNOLD, *On matrices depending on parameters*, Russian Math. Surveys, 26 (1971), pp. 29–43.
[2] T. BEELEN AND P. VAN DOOREN, *An improved algorithm for the computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 105 (1988), pp. 9–65.
[3] J. M. BERG AND H. G. KWATNY, *A canonical parameterization of the Kronecker form of a matrix pencil*, Automatica, 31 (1995), pp. 669–680.
[4] T. BRÖCKER AND L. LANDER, *Differential Germs and Catastrophes*, Cambridge University Press, London, UK, 1975.
[5] J. W. BRUCE AND P. J. GIBLIN, *Curves and Singularities*, Cambridge University Press, London, UK, 1991.
[6] S.-N. CHOW, C. LI, AND D. WANG, *Normal Forms and Bifurcation of Planar Vector Fields*, Cambridge University Press, London, UK, 1994.
[7] J. DEMMEL, *On structured singular values*, in Proc. of the 27th Conference on Decision and Control, Austin, TX, IEEE, 1988.
[8] J. DEMMEL AND A. EDELMAN, *The dimension of matrices (matrix pencils) with given Jordan (Kronecker) canonical forms*, Linear Algebra Appl., 230 (1995), pp. 61–87.
[9] J. DEMMEL AND B. KÅGSTRÖM, *Computing stable eigendecompositions of matrix pencils*, Linear Algebra Appl., 88/89 (1987), pp. 139–186.
[10] J. DEMMEL AND B. KÅGSTRÖM, *Accurate solutions of ill-posed problems in control theory*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 126–145.

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

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

[13] A. Edelman and H. Murakami, *Polynomial roots from companion matrix eigenvalues*, Math. Comp., 64 (1995), pp. 763–776.

[14] A. Edelman, E. Elmroth, and B. Kågström, *A Geometric Approach to Perturbation Theory of Matrices and Matrix Pencils. Part* II: *A Stratification-Enhanced Staircase Algorithm*, Report UMINF-96.13, Department of Computing Science, Umeå University, Umeå, Sweden, 1996; SIAM J. Matrix Anal. Appl., submitted.

[15] E. Elmroth and B. Kågström, *The set of* 2*-by-*3 *matrix pencils—Kronecker structures and their transitions under perturbations*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 1–34.

[16] F. Gantmacher, *The Theory of Matrices, Vol.* I *and* II, translation, Chelsea, NY, 1959.

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

[18] M. Golubitsky and V. Guillemin, *Stable Mappings and their Singularities*, Springer-Verlag, Berlin, New York, 1973.

[19] U. Helmke and D. Hinrichsen, *Canonical forms and orbit spaces of linear systems*, IMA J. Math. Control Inform., 3 (1986), pp. 167–184.

[20] D. Hinrichsen and D. Prätzel-Wolters, *A wild quiver in linear systems theory*, Linear Algebra Appl., 91 (1987), pp. 143–175.

[21] D. Hinrichsen and D. Prätzel-Wolters, *A Jordan canonical form for reachable linear systems*, Linear Algebra Appl., 122/123/124 (1989), pp. 489–524.

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

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

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

[25] Y.-C. Lu, *Singularity Theory and an Introduction to Catastrophe Theory*, Springer-Verlag, Berlin, New York, 1976.

[26] A. Pokrzywa, *On perturbations and the equivalence orbit of a matrix pencil*, Linear Algebra Appl., 82 (1986), pp. 99–121.

[27] D. Prätzel-Wolters, *Canonical forms for linear systems*, Linear Algebra Appl., 50 (1983), pp. 437–473.

[28] M. Shub and S. Smale, *Complexity of Bezout's theorem* II: *Volumes and probabilities*, in Computational Algebraic Geometry, F. Eyssette and A. Galligo, eds., Progress in Mathematics 109, Birkhauser, 1993, pp. 267–285.

[29] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[30] L. Stolovitch, *On the computation of a versal family of matrices*, Numer. Algorithms, 4 (1993), pp. 25–56.

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

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

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

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

[35] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford Science Publications, Oxford, UK, 1965.

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

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 3, pp. 693–705, July 1997

009

ON THE SHAPE OF THE SYMMETRIC, PERSYMMETRIC, AND SKEW-SYMMETRIC SOLUTION SET*

GÖTZ ALEFELD†, VLADIK KREINOVICH‡, AND GÜNTER MAYER§

Dedicated to Prof. Dr. Gerhard Maeß, Rostock, on the occasion of his 60*th birthday.*

Abstract. We present a characterization of the solution set S, the symmetric solution set S_{sym}, the persymmetric solution set S_{per}, and the skew-symmetric solution set S_{skew} of real linear systems $Ax = b$ with the $n \times n$ coefficient matrix A varying between a lower bound $\underline{A}$ and an upper bound $\overline{A}$, and with b similarly varying between $\underline{b}$, $\overline{b}$. We show that in each orthant the sets S_{sym}, S_{per}, and S_{skew} are, respectively, the intersection of S with sets, the boundaries of which are quadrics.

Key words. linear systems with perturbed input data, solution set of linear systems of equations, symmetric matrices, persymmetric matrices, skew-symmetric matrices, Oettli–Prager theorem, Fourier–Motzkin elimination, interval analysis

AMS subject classifications. 65G10, 65F05

PII. S0895479896297069

1. Introduction.

Let $[A]$ be an $n \times n$ matrix with compact intervals as entries, let $[b]$ be a vector with n interval components, and let E be the $n \times n$ permutation matrix with ones in the northeast–southwest diagonal and zeros elsewhere. The purpose of this paper is to characterize the solution sets

$$S := \{x \in \boldsymbol{R}^n \mid Ax = b,\ A \in [A],\ b \in [b]\}, \tag{1.1}$$

$$S_{sym} := \{x \in \boldsymbol{R}^n \mid Ax = b,\ A = A^T \in [A] = [A]^T,\ b \in [b]\}, \tag{1.2}$$

$$S_{per} := \{x \in \boldsymbol{R}^n \mid Ax = b,\ EA = (EA)^T \in E[A] = (E[A])^T,\ b \in [b]\}, \tag{1.3}$$

$$S_{skew} := \{x \in \boldsymbol{R}^n \mid Ax = b,\ A = -A^T \in [A] = ([a]_{ij}) = -[A]^T, \quad [a]_{ii} = 0 \text{ for } i = 1, \ldots, n,\ b \in [b]\} \tag{1.4}$$

by means of inequalities which show that in each fixed orthant O the solution set S is the intersection of finitely many half spaces, while $S_{sym} \cap O$, $S_{per} \cap O$, and $S_{skew} \cap O$ are the intersection of $S \cap O$ with finitely many sets, the boundaries of which are conic sections in $\boldsymbol{R}^n$. The characterization of $S \cap O$ was already given in [4], [5], [7], [11], [12], and others while the characterization of $S_{sym} \cap O$ in the two-dimensional case was derived in [4]. The technique there could not be transferred onto the general case in an obvious way. It was changed in [2], [3]. We will use here a different technique known as Fourier–Motzkin elimination, which is described, e.g., in [14].

*Received by the editors January 8, 1996; accepted for publication (in revised form) by N. J. Higham July 18, 1996.

http://www.siam.org/journals/simax/18-3/29706.html

†Institut für Angewandte Mathematik, Universität Karlsruhe, D-76128 Karlsruhe, Germany (goetz.alefeld@mathematik.uni-karlsruhe.de).

‡Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968 (vladik@cs.utep.edu).

§Fachbereich Mathematik, Universität Rostock, D-18051 Rostock, Germany (guenter.mayer@mathematik.uni-rostock.de).

Note that we require

$$(1.5)\quad \begin{cases} \text{no additional condition on } [A] \text{ in the case of } S, \\ [A] = [A]^T \text{ in the case of } S_{sym}, \\ E[A] = (E[A])^T \text{ in the case of } S_{per}, \\ [A] = ([a]_{ij}) = -[A]^T \text{ with } [a]_{ii} = 0,\ i = 1, \dots, n \text{ in the case of } S_{skew}. \end{cases}$$

The restrictions in (1.5) are not severe. If $[A] \neq [A]^T$ in the case of S_{sym}, e.g., and if $[B]$ denotes the largest interval matrix in $[A]$ such that $[B] = [B]^T$ holds, then the matrices in $[A]\backslash[B]$ do not influence S_{sym}. Therefore, instead of $[A]$ the matrix $[B]$ would play the crucial role in characterizing S_{sym}.

We emphasize that $[A]$ is allowed to contain singular real matrices. The restriction $[a]_{ii} = 0,\ i = 1, \dots, n$ in the case of S_{skew} stems from the fact that a skew-symmetric matrix $A = (a_{ij}) \in \boldsymbol{R}^{n\times n}$ is defined by $A = -A^T$ which implies $a_{ii} = 0$ for $i = 1, \dots, n$. We also recall that this matrix is singular if n is odd. This can be seen from $\det A = \det(-A^T) = \det(-A) = (-1)^n \det A$. The condition $EA = (EA)^T$ for S_{per} characterizes a persymmetric matrix which is defined to be symmetric with respect to the northeast–southwest diagonal; cf. [6], e.g.

The sets in (1.1)–(1.4) occur when dealing with linear systems of equations, the input data of which are afflicted with tolerances (cf. [1], [10], or [13], e.g.). This is the case when data $\check{A}$, $\check{b}$ are perturbed by errors caused, e.g., by measurements or by a conversion from decimal to binary digits on a computer. Assume that these errors are known to be bounded by some quantities $\Delta A \in \boldsymbol{R}^{n\times n}$ and $\Delta b \in \boldsymbol{R}^n$ with nonnegative entries. Then it seems reasonable to accept a vector $\tilde{x}$ as the "correct" solution of $\check{A}x = \check{b}$ if it is in fact the solution of a perturbed system $\tilde{A}x = \tilde{b}$ with

$$\tilde{A} \in [A] := [\check{A} - \Delta A, \check{A} + \Delta A], \quad \tilde{b} \in [b] := [\check{b} - \Delta b, \check{b} + \Delta b].$$

The characterization of all such $\tilde{x}$ led Oettli and Prager [11] to their famous equivalence

$$(1.6)\qquad x \in S \Longleftrightarrow |\,\check{b} - \check{A}x\,| \leq \Delta A|x| + \Delta b,$$

where $|v| := (|v_i|) \in \boldsymbol{R}^n$ for $v = (v_i) \in \boldsymbol{R}^n$. It relates the midpoint residual to the tolerances and to $|x|$ and was reformulated in [7] similarly as in the subsequent Theorem 3.4. Often $\check{A}$ belongs to a particular class of matrices with dependencies in their entries. Such a class is formed by symmetric matrices, persymmetric matrices, skew-symmetric matrices, and others. Therefore, it is reasonable to consider subsets of S for which the elements x are solutions of linear systems $Ax = b$ with *special* matrices A only. This leads to the problem discussed in this paper. Our results are formulated in terms of inequalities involving the bounds of $[A]$, $[b]$. They can easily be reformulated using the midpoints $\check{A}$, $\check{b}$ and the tolerances ΔA, Δb, although a compact form such as (1.6) is still missing.

We also mention that the sets S_{sym} and S_{skew} were already considered in [8] and [9]. There, bounds for the projections of these sets onto the coordinate axes were derived but no characterization of these sets were given.

We have arranged our paper as follows. In section 2 we list the notation which we will use throughout the paper; in section 3 we present the results. We close our paper with some examples in section 4 which illustrate the technique and the theory.

2. Preliminaries. By $\boldsymbol{R}^n$, $\boldsymbol{R}^{n\times n}$, $\boldsymbol{IR}$, $\boldsymbol{IR}^n$, and $\boldsymbol{IR}^{n\times n}$ we denote the set of real vectors with n components, the set of real $n \times n$ matrices, the set of intervals, the set of interval vectors with n components, and the set of $n \times n$ interval matrices, respectively. By "interval" we always mean a real compact interval. Interval vectors and interval matrices are vectors and matrices, respectively, with interval entries. We write intervals in brackets with the exception of degenerate intervals (so-called *point intervals*), which we identify with the element being contained, and we proceed similarly with interval vectors and interval matrices. We write $[A] = [\underline{A}, \overline{A}] = ([a]_{ij}) = ([\underline{a}_{ij}, \overline{a}_{ij}]) \in \boldsymbol{IR}^{n\times n}$ simultaneously, without further reference, and we use an analogous notation for intervals and interval vectors. By $[A]^T$ we mean the transposed matrix of $[A]$. We mention that $[A] = [A]^T$ is equivalent to $\underline{A} = \underline{A}^T$ and $\overline{A} = \overline{A}^T$ and that $[A] = -[A]^T$ is equivalent to $\underline{A} = -\overline{A}^T$ and $\overline{A} = -\underline{A}^T$. Therefore, if an interval matrix $[A]$ fulfills the condition $[A] = -[A]^T$, its midpoint matrix $\check{A} := \frac{1}{2}(\underline{A} + \overline{A})$ satisfies $\check{A} = -\check{A}^T$; i.e., $\check{A}$ is skew-symmetric. We call an $n \times n$ interval matrix *singular* if it contains at least one singular real matrix; otherwise, we call it *regular*. For computations with interval quantities we refer to [1] or [10].

By O we denote any closed orthant of $\boldsymbol{R}^n$. To distinguish among the sets S, S_{sym}, S_{per}, and S_{skew} we call S_{sym} the *symmetric solution set*, S_{per} the *persymmetric solution set*, and S_{skew} the *skew-symmetric solution set*.

3. Results. We start this section with a topological result which for S and S_{sym} is already known (see [4]).

THEOREM 3.1. *Let* $[A] \in \boldsymbol{IR}^{n\times n}$ *be regular and satisfy* (1.5).

(a) *Each of the sets* S_{sym}, S_{per}, S_{skew}, $S \cap O$, $S_{sym} \cap O$, $S_{per} \cap O$, *and* $S_{skew} \cap O$ *is compact.*

(b) *Each of the sets* S, S_{sym}, S_{per}, S_{skew}, *and* $S \cap O$ *is connected;* $S \cap O$ *is convex.*

Proof. First, we prove the assertions for S_{skew}. Let $A = -A^T \in [A]$ and interpret $x = A^{-1}b$ as a function f of the $\frac{n(n-1)}{2}$ variables a_{ij}, $1 \le i < j \le n$ and the n variables b_i, $1 \le i \le n$. This function is continuous. Since $[a]_{ij}$, $[b]_i$ are connected and compact the same holds for the range S_{skew} of f.

The compactness of the intersection $S_{skew} \cap O$ follows from S_{skew} being compact and from O being closed.

In the cases of S, S_{sym}, and S_{per} one proves the assertions by similar arguments.

The convexity of $S \cap O$ results from the fact that this set can be expressed as the intersection of finitely many half spaces (cf. [11] or the subsequent Theorem 3.4, e.g.). □

Remark. If $[A]$ is singular but contains no singular *symmetric* matrix the proof of Theorem 3.1 shows that S_{sym} remains compact and connected and that $S_{sym} \cap O$ remains compact. An analogous statement holds for S_{per}, S_{skew}, $S_{per} \cap O$, and $S_{skew} \cap O$. For singular $[A]$ the solution set S, however, is empty or unbounded since the kernel of each singular matrix $A \in [A]$ is unbounded. Due to singularity, the function f with $f(A, b) := A^{-1}b$ is certainly not defined on $[A] \times [b]$. This already indicates that the assertions of Theorem 3.1 may be wrong in the singular case. As an illustration we consider the example

$$[A] := \begin{pmatrix} 0 & [-1,1] \\ [-1,1] & 0 \end{pmatrix}, \qquad [b] := \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Any real matrix $A \in [A]$ can be represented by

$$A = \begin{pmatrix} 0 & \alpha \\ \beta & 0 \end{pmatrix}$$

with α, $\beta \in [-1, 1]$. Hence A is regular with

$$A^{-1} = \begin{pmatrix} 0 & \beta^{-1} \\ \alpha^{-1} & 0 \end{pmatrix}$$

provided that $\alpha\beta \neq 0$. We obtain

$$\begin{aligned} S &= \{\, (\gamma, \delta)^T \mid \gamma \in \boldsymbol{R},\ -\infty < \delta \leq -1 \text{ or } 1 \leq \delta < \infty \,\}, \\ S_{sym} = S_{skew} &= \{\, (0, \delta)^T \mid\ -\infty < \delta \leq -1 \text{ or } 1 \leq \delta < \infty \,\}, \end{aligned}$$

which shows that neither S nor S_{sym} nor S_{skew} is compact or connected in this case. □

Our next theorem characterizes S_{skew} by a set of inequalities. Its proof starts with

$$x \in S_{skew} \Longleftrightarrow \underline{b} \leq Ax \leq \overline{b},\ A = -A^T \in [A], \tag{3.1}$$

transforms the inequalities in a suitable way by introducing new variables z_{ij}, and continues by applying the Fourier–Motzkin elimination (see [14], e.g.) to replace the entries of A by their bounds $\underline{a}_{ij}$ and $\overline{a}_{ij}$, respectively.

THEOREM 3.2. *Let $[A] = -[A]^T \in \boldsymbol{IR}^{n\times n}$ with $[a]_{ii} = 0,\ i = 1, \ldots, n$, and let $[b] \in \boldsymbol{IR}^n$. Then for any orthant $O \subseteq \boldsymbol{R}^n$ the set $S_{skew} \cap O$ can be represented as an intersection of finitely many closed sets, the boundaries of which are quadrics or hyperplanes. The inequalities characterizing these hyperplanes and quadrics can be derived from $\underline{b} \leq Ax \leq \overline{b},\ A = -A^T \in [A],\ x \in O$ by means of the Fourier–Motzkin elimination.*

Proof. Step 1. Let (3.1) hold, fix an orthant O, and define

$$\begin{aligned} a_{ij}^- &:= \begin{cases} \underline{a}_{ij} & \text{if } x_i x_j \geq 0, \\ \overline{a}_{ij} & \text{if } x_i x_j < 0, \end{cases} & a_{ij}^+ &:= \begin{cases} \overline{a}_{ij} & \text{if } x_i x_j \geq 0, \\ \underline{a}_{ij} & \text{if } x_i x_j < 0, \end{cases} \\ b_i^- &:= \begin{cases} \underline{b}_i & \text{if } x_i \geq 0, \\ \overline{b}_i & \text{if } x_i < 0, \end{cases} & b_i^+ &:= \begin{cases} \overline{b}_i & \text{if } x_i \geq 0, \\ \underline{b}_i & \text{if } x_i < 0. \end{cases} \end{aligned} \tag{3.2}$$

Note that the values of a_{ij}^-, a_{ij}^+, b_i^-, b_i^+ are constant as long as x remains in the same orthant and that they satisfy $a_{ij}^- = -a_{ji}^+$ and $a_{ii}^- = a_{ii}^+ = 0$. We first will see that (3.1) is equivalent to

$$\begin{cases} x \in S \quad \wedge \quad \exists\, z_{ij} \in \boldsymbol{R} \text{ such that} \\ \qquad \begin{cases} a_{ij}^- x_i x_j \leq z_{ij} \leq a_{ij}^+ x_i x_j, & i, j = 1, \ldots, n,\ i < j, \\ z_{ij} = -z_{ji}, & i, j = 1, \ldots, n, \\ b_i^- x_i \leq \sum_{j=1}^n z_{ij} \leq b_i^+ x_i, & i = 1, \ldots, n. \end{cases} \end{cases} \tag{3.3}$$

Setting $z_{ij} := a_{ij} x_i x_j$ immediately shows that "(3.1) $\Rightarrow$ (3.3)." To prove the converse we will construct $A \in \boldsymbol{R}^{n\times n}$ such that $A = -A^T \in [A]$ and $Ax \in [b]$. Consider a fixed index pair i_0, j_0 and define $a_{i_0 j_0}$ according to the following procedure.

Case 1: $x_{i_0} = 0$. Since $x \in S$ by (3.3), there are real numbers $a^*_{i_0 j}$ for $j = 1, \ldots, n$ such that

$$\underline{a}_{i_0 j} \le a^*_{i_0 j} \le \overline{a}_{i_0 j} \tag{3.4}$$

and

$$\underline{b}_{i_0} \le \sum_{j=1}^{n} a^*_{i_0 j} x_j \le \overline{b}_{i_0}. \tag{3.5}$$

If $x_{j_0} \neq 0$ then $a_{i_0 j_0} := a^*_{i_0 j_0} =: -a_{j_0 i_0}$; if $x_{j_0} = 0$ then $a_{i_0 j_0} := \check{a}_{i_0 j_0}$ with $\check{a}_{i_0 j_0}$ being the corresponding entry of the skew-symmetric midpoint matrix $\check{A} \in [A]$.

Case 2: $x_{i_0} \neq 0$. If $x_{j_0} \neq 0$ then $a_{i_0 j_0} := \frac{z_{i_0 j_0}}{x_{i_0} x_{j_0}}$; if $x_{j_0} = 0$ then $a_{i_0 j_0}$ is already defined by the preceding case when the roles of i_0 and j_0 are exchanged.

If one lets i_0 run from 1 to n and if for each fixed i_0 the second index in $z_{i_0 j_0}$ runs from 1 to n then by the procedure above a skew-symmetric matrix $A \in [A]$ is constructed which satisfies (3.1). Note that in Case 1 of our procedure there may occur several choices for the entries $a^*_{i_0 j}$ such that (3.4) and (3.5) are valid. It is obvious that in this case for a fixed i_0 the entries of one and the same double inequality (3.5) must be chosen for those $j_0 = 1, \ldots, n$ for which $x_{j_0} \neq 0$. Together with the last double inequality in (3.3), this guarantees $\underline{b}_i \le \sum_{j=1}^{n} a_{ij} x_j \le \overline{b}_i$.

The condition "$x \in S$" in (3.3) is necessary, as the example $A := 0 \in \boldsymbol{R}^{1\times 1}$, $b := 1 \in \boldsymbol{R}$ shows. Here, $x = 0 \in \boldsymbol{R}$ is clearly not in $S \supseteq S_{skew}$, but the remaining conditions of (3.3) are fulfilled for $z_{11} = 0$.

Step 2. By $z_{ii} = -z_{ii}$ we obtain $z_{ii} = 0$. Therefore, we omit z_{ii} in (3.3). We now apply the Fourier–Motzkin elimination to (3.3). We illustrate this process by eliminating z_{12}. To this end we replace z_{ij} by $-z_{ji}$ for all $i > j$ in the inequalities of (3.3). We rewrite these inequalities and change their order by forming three groups: the inequalities of the first group have the form $\cdots \le z_{12}$ with z_{12}-free left-hand side, the inequalities of the second group read $z_{12} \le \cdots$ with z_{12}-free right-hand side, and the inequalities of the third group do not contain z_{12}. Since the maximum over all left-hand sides of the inequalities of the first group is less than or equal to the minimum over all right-hand sides of the inequalities of the second group, these inequalities are equivalent to requiring that *each* left-hand side of the first group be less than or equal to *each* right-hand side of the second group while keeping all inequalities of the third group. Omitting trivial inequalities, (3.3) is equivalent to

$$\left\{\begin{array}{lll} x \in S & \wedge & \exists\, z_{ij} \in \boldsymbol{R} \text{ such that} \\ & & \left\{\begin{array}{l} a^-_{12} x_1 x_2 \le b^+_1 x_1 - \sum_{j=3}^{n} z_{1j}, \\ a^-_{12} x_1 x_2 \le -b^-_2 x_2 + \sum_{j=3}^{n} z_{2j}, \\ b^-_1 x_1 - \sum_{j=3}^{n} z_{1j} \le a^+_{12} x_1 x_2, \\ b^-_1 x_1 - \sum_{j=3}^{n} z_{1j} \le -b^-_2 x_2 + \sum_{j=3}^{n} z_{2j}, \\ -b^+_2 x_2 + \sum_{j=3}^{n} z_{2j} \le a^+_{12} x_1 x_2, \\ -b^+_2 x_2 + \sum_{j=3}^{n} z_{2j} \le b^+_1 x_1 - \sum_{j=3}^{n} z_{1j}, \\ \text{remaining (in)equalities of (3.3)}, \end{array}\right. \end{array}\right. \tag{3.6}$$

where z_{12} and z_{21} no longer occur. This process of eliminating z_{ij} can be continued until we end up with a set of final inequalities which (together with $x \in S \cap O$) is equivalent to $x \in S_{skew} \cap O$ and which contains no variable z_{ij}. This proves the theorem. ∎

At the end of the elimination process, there are two special inequalities for each $i \in \{1, \ldots, n\}$ which can be divided by $x_i \neq 0$ such that no fractions occur. For example, if the first inequality of (3.6) is combined successively with the inequalities $a_{1j}^- x_1 x_j \leq z_{1j}$ one obtains the final inequality $\sum_{j=2}^n a_{1j}^- x_1 x_j \leq b_1^+ x_1$. Since $a_{11}^- = a_{11}^+ = 0$ it can be supplemented to $\sum_{j=1}^n a_{1j}^- x_1 x_j \leq b_1^+ x_1$, which reduces to

$$\sum_{j=1}^n a_{1j}^- x_j \leq \overline{b}_1 \quad \text{if } x_1 > 0 \quad \text{and} \quad \sum_{j=1}^n a_{1j}^- x_j \geq \underline{b}_1 \quad \text{if } x_1 < 0. \tag{3.7}$$

From the third inequality of (3.6) one similarly obtains

$$\sum_{j=1}^n a_{1j}^+ x_j \geq \underline{b}_1 \quad \text{if } x_1 > 0 \quad \text{and} \quad \sum_{j=1}^n a_{1j}^+ x_j \leq \overline{b}_1 \quad \text{if } x_1 < 0. \tag{3.8}$$

With

$$\hat{a}_{ij}^- := \begin{cases} \underline{a}_{ij} & \text{if } x_j \geq 0, \\ \overline{a}_{ij} & \text{if } x_j < 0, \end{cases} \qquad \hat{a}_{ij}^+ := \begin{cases} \overline{a}_{ij} & \text{if } x_j \geq 0, \\ \underline{a}_{ij} & \text{if } x_j < 0, \end{cases} \tag{3.9}$$

the four inequalities in (3.7) and (3.8) can be summarized to

$$\sum_{j=1}^n \hat{a}_{1j}^- x_j \leq \overline{b}_1 \quad \text{and} \quad \sum_{j=1}^n \hat{a}_{1j}^+ x_j \geq \underline{b}_1,$$

provided that $x_1 \neq 0$. Repeating the arguments, one finally gets

$$\left.\begin{array}{l} \sum_{j=1}^n \hat{a}_{ij}^- x_j \leq \overline{b}_i, \\ \sum_{j=1}^n \hat{a}_{ij}^+ x_j \geq \underline{b}_i, \end{array}\right\} \quad i = 1, \ldots, n \tag{3.10}$$

if no component of x equals 0. These inequalities are just those which characterize S and which are known as the Oettli–Prager theorem (cf. [11]), which we restate as Theorem 3.4. They can either be omitted in the list of inequalities if "$x \in S$" remains there as in (3.6), or "$x \in S$" can be cancelled when (3.10) is used instead. This last remark also holds if some of the components of x are zero.

We also note that the number $n_\#$ of final inequalities for $S_{skew} \cap O$ seems to be double exponential. Thus we could show that $n_\#$ is roughly bounded by $8 \cdot \left(\frac{3}{2}\right)^{2^{\kappa+1}}$ with $\kappa := \frac{n(n+1)}{2}$. Since the arguments are a little bit clumsy and the proof is lengthy we will skip it.

The same technique for S_{skew} can also be applied to construct a set of inequalities which characterize S_{sym} provided that $[A] = [A]^T$. To get the equivalence to "$x \in S_{sym}$" one must replace the equality in (3.1) by $A = A^T$, and one uses $z_{ij} = z_{ji}$ in (3.3) instead of $z_{ij} = -z_{ji}$. Analogously to Theorem 3.2, we get the following theorem.

THEOREM 3.3. *Let $[A] = [A]^T \in \boldsymbol{IR}^{n \times n}$ and let $[b] \in \boldsymbol{IR}^n$. Then for any orthant $O \subseteq \boldsymbol{R}^n$ the set $S_{sym} \cap O$ can be represented as an intersection of finitely many closed sets, the boundaries of which are quadrics or hyperplanes. The inequalities characterizing these hyperplanes and quadrics can be derived from the elimination process described above or they are of the form $x_i = 0$.* □

Theorem 3.3 can analogously be formulated for S_{per} since $Ax = b \Longleftrightarrow EAx = Eb$, whence S_{per} for A equals S_{sym} for $EA = (EA)^T$.

The solution set for other classes of special matrices such as Hankel or Toeplitz matrices shows particularities which essentially differ from those which we have presented up to now. Thus, the inequalities need no longer remain the same in a fixed orthant and they cannot be treated by means of the particular variables z_{ij}. Work in this respect is in progress.

Inequalities (3.10) can also be obtained with the technique above if one starts with

$$x \in S \tag{3.11}$$

instead of $x \in S_{skew}$. The conditions corresponding to (3.3) then read

$$\exists\, z_{ij} \in \boldsymbol{R} \text{ such that } \begin{cases} \hat{a}_{ij}^- x_j \le z_{ij} \le \hat{a}_{ij}^+ x_j, & i,j = 1,\dots,n, \\ \underline{b}_i \le \sum_{j=1}^n z_{ij} \le \overline{b}_i, & i = 1,\dots,n \end{cases} \tag{3.12}$$

with $\hat{a}_{ij}^-$, $\hat{a}_{ij}^+$ from (3.9). To prove the implication "(3.12) $\Rightarrow$ (3.11)" set $a_{ij} = \frac{z_{ij}}{x_j}$ if $x_j \neq 0$. If $x_j = 0$ then any element from $[a]_{ij}$ can be used to construct a matrix A such that $Ax \in [b]$ holds. It is easy to see that one ends up with inequalities (3.10) if one performs the elimination process as above, starting with (3.12).

For completeness we state the result in a separate theorem.

THEOREM 3.4 (Oettli–Prager theorem [11]). *Let $[A] \in \boldsymbol{IR}^{n\times n}$ and let $[b] \in \boldsymbol{IR}^n$. Then for any orthant $O \subseteq \boldsymbol{R}^n$ the set $S \cap O$ can be represented as the intersection of closed half spaces. These half spaces are given by*

$$\left.\begin{array}{l} \sum_{j=1}^n \hat{a}_{ij}^- x_j \le \overline{b}_i, \\ \sum_{j=1}^n \hat{a}_{ij}^+ x_j \ge \underline{b}_i, \end{array}\right\} \quad i = 1,\dots,n \tag{3.13}$$

or

$$x_i \le 0 \quad \text{or} \quad x_i \ge 0, \tag{3.14}$$

where the inequalities in (3.14) *are used to characterize the orthant O and where $\hat{a}_{ij}^-$, $\hat{a}_{ij}^+$ are defined in* (3.9). ∎

4. Examples. In this section we present several examples to illustrate the results of section 3. In particular, we construct the inequalities for characterizing S, S_{sym}, S_{per}, and S_{skew}.

In our first example we consider 2×2 interval matrices.

Example 4.1.

(a) Let $[A] \in \boldsymbol{IR}^{2\times 2}$, $[b] \in \boldsymbol{IR}^2$. Then S is characterized according to (3.13) by the inequalities

$$\begin{cases} \hat{a}_{11}^- x_1 + \hat{a}_{12}^- x_2 \le \overline{b}_1, & \hat{a}_{11}^+ x_1 + \hat{a}_{12}^+ x_2 \ge \underline{b}_1, \\ \hat{a}_{21}^- x_1 + \hat{a}_{22}^- x_2 \le \overline{b}_2, & \hat{a}_{21}^+ x_1 + \hat{a}_{22}^+ x_2 \ge \underline{b}_2 \end{cases} \tag{4.1}$$

with the coefficients according to (3.9).

(b) Let $[A] = [A]^T$ hold. The symmetric solution set S_{sym} is described by the four inequalities in (4.1) supplemented by the two inequalities

$$\begin{cases} b_1^- x_1 - b_2^+ x_2 - a_{11}^+ x_1^2 + a_{22}^- x_2^2 \le 0, \\ -b_1^+ x_1 + b_2^- x_2 + a_{11}^- x_1^2 - a_{22}^+ x_2^2 \le 0 \end{cases} \tag{4.2}$$

with the coefficients from (3.2). These inequalities show that the boundary of S_{sym} can already be curvilinear in the 2×2 case.

(c) Let $E[A] = (E[A])^T$ hold. The persymmetric solution set S_{per} is described by the four inequalities in (4.1) supplemented by the two inequalities in (4.2) if one redefines $a_{ii}^{\pm}$, $b_i^{\pm}$ appropriately.

(d) Let $[A] = -[A]^T$ hold with $[a]_{ii} = 0$ for $i = 1, 2$. The skew-symmetric solution set S_{skew} is given by the four inequalities in (4.1) with $\hat{a}_{ii}^- = \hat{a}_{ii}^+ = 0$ in addition to the two inequalities

$$b_1^- x_1 \le -b_2^- x_2, \qquad -b_2^+ x_2 \le b_1^+ x_1, \tag{4.3}$$

which follow directly from (3.6) taking into account $z_{11} = z_{22} = 0$. The skew-symmetric solution set in $\boldsymbol{R}^2$ is apparently bounded by a polygon; i.e., its boundary is formed by straight lines. Taking into account $\hat{a}_{ii}^- = \hat{a}_{ii}^+ = 0$, one sees immediately from (4.1) that the solution set S is an interval vector. This is not always the case for S_{skew}. For example, choose $[b] := (1, 1)^T$ and $[a]_{12} := [0.25, 1]$. Then any skew-symmetric element A of $[A]$ can be written in the form

$$A = \alpha \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = -\alpha^2 A^{-1} \qquad \text{with } 0.25 \le \alpha \le 1.$$

Hence $S_{skew} = \{\, \beta(-1, 1)^T \mid 1 \le \beta \le 4\}$; i.e., S_{skew} is the straight line in the plane between the points $(-1, 1)$ and $(-4, 4)$. The corresponding solution set S, however, is given by

$$S = \{\, (-\beta, \gamma)^T \mid 1 \le \beta, \gamma \le 4 \,\} = ([-4, -1], [1, 4])^T. \qquad \square$$

In our second example we consider 3×3 tridiagonal interval matrices.

Example 4.2.

(a) Let $[A] \in \boldsymbol{IR}^{3\times 3}$ with $[a]_{13} = [a]_{31} := 0$, and let $[b] \in \boldsymbol{R}^3$. Then S is characterized by the inequalities

$$\left\{ \begin{array}{rr} \hat{a}_{11}^- x_1 + \hat{a}_{12}^- x_2 \le \overline{b}_1, & \hat{a}_{11}^+ x_1 + \hat{a}_{12}^+ x_2 \ge \underline{b}_1, \\ \hat{a}_{21}^- x_1 + \hat{a}_{22}^- x_2 + \hat{a}_{23}^- x_3 \le \overline{b}_2, & \hat{a}_{21}^+ x_1 + \hat{a}_{22}^+ x_2 + \hat{a}_{23}^+ x_3 \ge \underline{b}_2, \\ \hat{a}_{32}^- x_2 + \hat{a}_{33}^- x_3 \le \overline{b}_3, & \hat{a}_{32}^+ x_2 + \hat{a}_{33}^+ x_3 \ge \underline{b}_3, \end{array} \right. \tag{4.4}$$

where the coefficients are again given by (3.9).

(b) For tridiagonal 3×3 matrices $[A] = [A]^T$ the symmetric solution set S_{sym} is characterized by the six inequalities in (4.4) and by the four additional inequalities

$$\left\{ \begin{array}{r} +b_1^- x_1 - b_2^+ x_2 - a_{11}^+ x_1^2 + a_{22}^- x_2^2 + a_{23}^- x_2 x_3 \le 0, \\ +b_1^- x_1 - b_2^+ x_2 + b_3^- x_3 - a_{11}^+ x_1^2 + a_{22}^- x_2^2 - a_{33}^+ x_3^2 \le 0, \\ +b_1^- x_1 - (+b_2^+ - b_2^-) x_2 - a_{11}^+ x_1^2 - a_{12}^+ x_1 x_2 - (+a_{22}^+ - a_{22}^-) x_2^2 \le 0, \\ +b_2^- x_2 - b_3^+ x_3 - a_{12}^+ x_1 x_2 - a_{22}^+ x_2^2 + a_{33}^- x_3^2 \le 0 \end{array} \right. \tag{4.5}$$

together with their four counterparts, which one gets by replacing each minus sign by a plus sign, and vice versa (also in the superscripts). The coefficients of (4.5) are defined in (3.2). Note that the information of the third inequality in (4.5) is contained in that of the first row of (4.4) if $[b]_2$ and $[a]_{22}$ are point intervals.

Without proof we mention that the number of inequalities for S_{sym} increases to 44 for a dense 3×3 system.

(c) The skew-symmetric solution set S_{skew} is characterized by (4.4) with $\hat{a}_{ii}^- = \hat{a}_{ii}^+ = 0$ for $i = 1, 2, 3$ and by the inequalities

$$\left\{\begin{array}{rl} -b_1^+ x_1 - b_2^+ x_2 + a_{23}^- x_2 x_3 \le 0, \\ +b_1^- x_1 + b_2^- x_2 + b_3^- x_3 \le 0, \\ +b_1^- x_1 - (+b_2^+ - b_2^-) x_2 - a_{12}^+ x_1 x_2 \le 0, \\ -b_2^+ x_2 - b_3^+ x_3 - a_{12}^+ x_1 x_2 \le 0 \end{array}\right. \tag{4.6}$$

together with their four counterparts, which are defined analogously as for S_{sym}. The inequalities in (4.6) look similar to those in (4.5) when taking into account $[a]_{ii} = 0$ for $i = 1, 2, 3$. Again, the third inequality in (4.6) equals the first one in (4.4) if $[b_2]$ is a point interval. Note also that according to section 1 each skew-symmetric matrix from $\boldsymbol{R}^{3\times 3}$ is singular! □

In our third example we describe S and S_{skew} in two different ways, a direct way (feasible since there is only one nontrivial pair of intervals) and a second way where we will apply the results of Example 4.2.

Example 4.3. Let

$$[A] := \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & [0.5, 1] \\ 0 & [-1, -0.5] & 0 \end{pmatrix}, \qquad [b] := \begin{pmatrix} [0, 2] \\ 0 \\ -1 \end{pmatrix}.$$

Then $[A] = -[A]^T$ with $[a]_{ii} = 0$, $i = 1, 2, 3$. Each $A \in [A]$, $b \in [b]$ can be represented as

$$A = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & \alpha \\ 0 & -\beta & 0 \end{pmatrix}, \qquad b = \begin{pmatrix} \gamma \\ 0 \\ -1 \end{pmatrix}$$

with $\alpha, \beta \in [0.5, 1]$, $\gamma \in [0, 2]$. The linear system $Ax = b$ then reads

$$x_2 = \gamma, \tag{4.7}$$

$$-x_1 + \alpha x_3 = 0, \tag{4.8}$$

$$-\beta x_2 = -1. \tag{4.9}$$

(a) We first want to describe the solution set S. Equations (4.7) and (4.8) show that $x_2 \ge 0$ and $\operatorname{sign}(x_1 x_3) \ge 0$. This means that only the first orthant O_1 and the sixth orthant O_6 can contain elements of S, where O_1 is characterized by $x_i \ge 0$, $i = 1, 2, 3$, and where O_6 is given by $x_1 \le 0$, $x_2 \ge 0$, $x_3 \le 0$. By the first and the third equation the system (4.7)–(4.9) is solvable if and only if $\beta\gamma = 1$. This is possible for any $\beta \in [0.5, 1]$ since $\gamma = \beta^{-1} \in [1, 2] \subseteq [0, 2]$. The solution can be rewritten as

$$x_1 = \alpha x_3, \quad x_2 = \beta^{-1}, \quad x_3 \in \boldsymbol{R}. \tag{4.10}$$

For each fixed α, $\beta \in [0.5, 1]$ these equations represent, of course, a straight line which lies in the plane $x_2 = \beta^{-1} \in [1, 2]$ and which crosses the x_2-axis at $(0, \beta^{-1}, 0)$. For each fixed $\beta \in [0.5, 1]$ one thus gets a (double) sector in $O_1 \cup O_6$ which is bounded by the straight lines $x_1 = 0.5x_3$ and $x_1 = x_3$ while $x_2 = \beta^{-1}$. Varying β results in two wedges, the cutting edges of which have length 1 and meet at the x_2-axis from $(0, 1, 0)$ to $(0, 2, 0)$.

(b) To characterize S_{skew} let $\alpha = \beta$. From (4.10) we then obtain $x_1 x_2 = x_3$ with $x_2 \in [1, 2]$, i.e., S_{skew} is the intersection of S with the hyperbolical paraboloid

$x_3 = x_1 x_2$ which transforms to $y_3 = y_1^2 - y_2^2$ via $x_1 = y_1 + y_2$, $x_2 = y_1 - y_2$, $x_3 = y_3$. In particular, the boundary of S_{skew} is curvilinear. Figure 1 shows $S \cap O_1$ and $S_{skew} \cap O_1$. The intersections $S \cap O_6$ and $S_{skew} \cap O_6$ are obtained by rotating the two sets around the x_2-axis by an amount of 180° degrees.

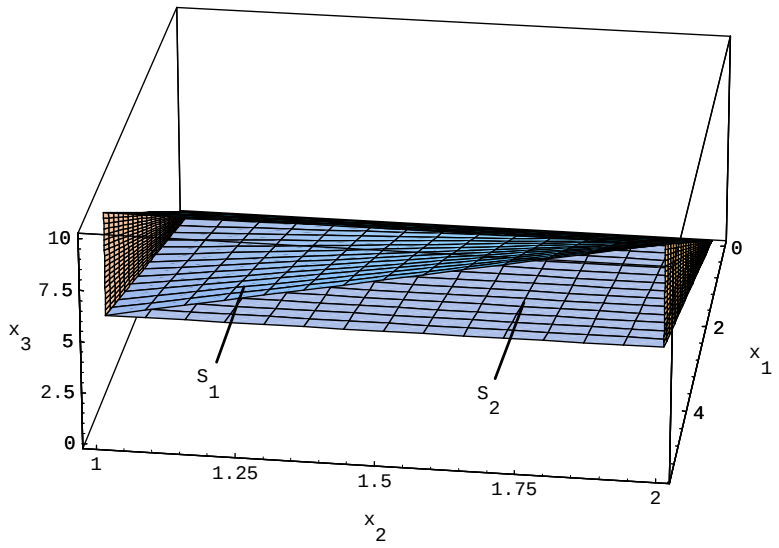


FIG. 4.1. *The shape of the solution sets* $S_1 := S \cap O_1$, $S_2 := S_{skew} \cap O_1$ *in Example* 4.3.

(c) We now want to describe S and S_{skew} in a second way, namely, by the inequalities resulting from (4.4) and (4.6). For simplicity we use $S \subseteq O_1 \cup O_6$, which yields $a_{23}^- = 0.5 = -a_{32}^+$, $a_{23}^+ = 1 = -a_{32}^-$. Inequalities (4.4) can then be written in the form

$$0 \le x_2 \le 2, \tag{4.11}$$

$$0.5x_3 \le x_1 \le x_3, \tag{4.12}$$

$$1 \le x_2 \le 2 \tag{4.13}$$

if $(x_1, x_2, x_3) \in O_1$. In O_6 inequality (4.12) must be replaced by $x_3 \le x_1 \le 0.5x_3$. Since (4.13) is more restrictive than (4.11) we can omit (4.11). Thus S is characterized by (4.12) and (4.13).

Inequalities (4.6) and their counterparts yield to

$$b_1^- x_1 \le x_2 x_3 \le 2b_1^+ x_1, \tag{4.14}$$

$$b_1^- x_1 \le x_3 \le b_1^+ x_1, \tag{4.15}$$

$$b_1^- x_1 \le x_1 x_2 \le b_1^+ x_1, \tag{4.16}$$

$$x_3 = x_1 x_2 \tag{4.17}$$

in O_1; in O_6 inequality (4.14) must be exchanged by $2b_1^- x_1 \le x_2 x_3 \le b_1^+ x_1$. Dividing (4.16) by x_1 implies (4.11). Hence (4.16) can be omitted. Since (4.15) is identical with (4.16) if (4.17) is used, we can skip (4.15) too. Replacing x_3 in (4.14) by (4.17) and dividing by x_1 yields to $0 = \underline{b}_1 \le x_2^2 \le 2\overline{b}_1 \le 4$, which again is fulfilled if (4.11) holds. Therefore, the inequalities for S_{skew} reduce in O_1 to

$$\begin{aligned} 1 \le x_2 \le 2, \\ x_1 \le x_3 \le 2x_1, \\ x_3 = x_1 x_2, \end{aligned}$$

which is equivalent to (4.10) when taking into account $\alpha = \beta \in [0.5, 1]$. The same holds in O_6 if the second double inequality is replaced by $2x_1 \le x_3 \le x_1$. ∎

In our last example we consider a 2×2 interval matrix $[A]$ which satisfies $[A] = [A]^T$.

Example 4.4. Let

$$[A] := \begin{pmatrix} 1 & [0,1] \\ [0,1] & [-4,-1] \end{pmatrix}, \qquad [b] := \begin{pmatrix} [0,2] \\ [0,2] \end{pmatrix}.$$

Then $[A] = [A]^T$ with

$$A = \begin{pmatrix} 1 & \alpha \\ \beta & -\gamma \end{pmatrix} \in [A] \quad \Longrightarrow \quad A^{-1} = \frac{1}{\gamma + \alpha\beta} \begin{pmatrix} \gamma & \alpha \\ \beta & -1 \end{pmatrix}$$

with $\alpha, \beta \in [0,1]$, $\gamma \in [1,4]$. Since $\underline{b} \ge 0$ the first component of $A^{-1}b$ is nonnegative for all $b \in [b]$. Therefore, S is completely contained in the union $O_1 \cup O_4$ of the first and the fourth quadrants.

We first consider $S \cap O_1$. According to (4.1) we get the inequalities

$$x_1 \le 2, \quad x_2 \ge -0.5, \quad x_2 \ge -x_1, \quad x_1 \ge x_2. \tag{4.18}$$

This means that $S \cap O_1$ is the triangle with the corners $(0,0)$, $(2,0)$, and $(2,2)$.

The corresponding inequalities for $S \cap O_4$ are given by

$$x_1 \ge 0, \quad x_2 \ge -2, \quad x_2 \le 2 - x_1, \quad x_2 \le 0.25\, x_1. \tag{4.19}$$

They describe a quadrangle with the corners $(0,0)$, $(0,-2)$, $(4,-2)$, and $(2,0)$.

To describe $S_{sym} \cap O_1$ we need inequalities (4.18) and the two inequalities from (4.2), which can be transform to

$$4x_1^2 + (4x_2 + 1)^2 \ge 1, \qquad (x_1 - 1)^2 + x_2^2 \le 1. \tag{4.20}$$

The first inequality of (4.20) describes an ellipse and its exterior. Since the ellipse lies completely in the lower half plane the first inequality of (4.20) is no restriction for $S_{sym} \cap O_1$. The second inequality describes a closed disc D_1 with center $(1,0)$ and radius 1. The boundary of the intersection with $S \cap O_1$ is formed by the straight line from $(0,0)$ to $(1,1)$, the part of the circle ∂D_1 from $(1,1)$ to $(2,0)$, and the part of the x_1-axis from $(2,0)$ back to $(0,0)$.

The inequalities in (4.19) together with the two inequalities

$$x_1^2 + 4x_2^2 \ge 0, \qquad (x_1 - 1)^2 + (x_2 + 1)^2 \le 2 \tag{4.21}$$

characterize $S_{sym} \cap O_4$. The first inequality in (4.21) is always true. The second inequality describes a disc D_2 with center $(1,-1)$ and radius $\sqrt{2}$. The boundary of its intersection with $S \cap O_4$ is formed by the straight lines from $(0,0)$ to $(0,-2)$, from $(0,-2)$ to $(2,-2)$, and from $(2,0)$ to $(0,0)$, and by the part of the circle ∂D_2 from $(2,-2)$ to $(2,0)$. The situation is illustrated by Figure 2. ∎

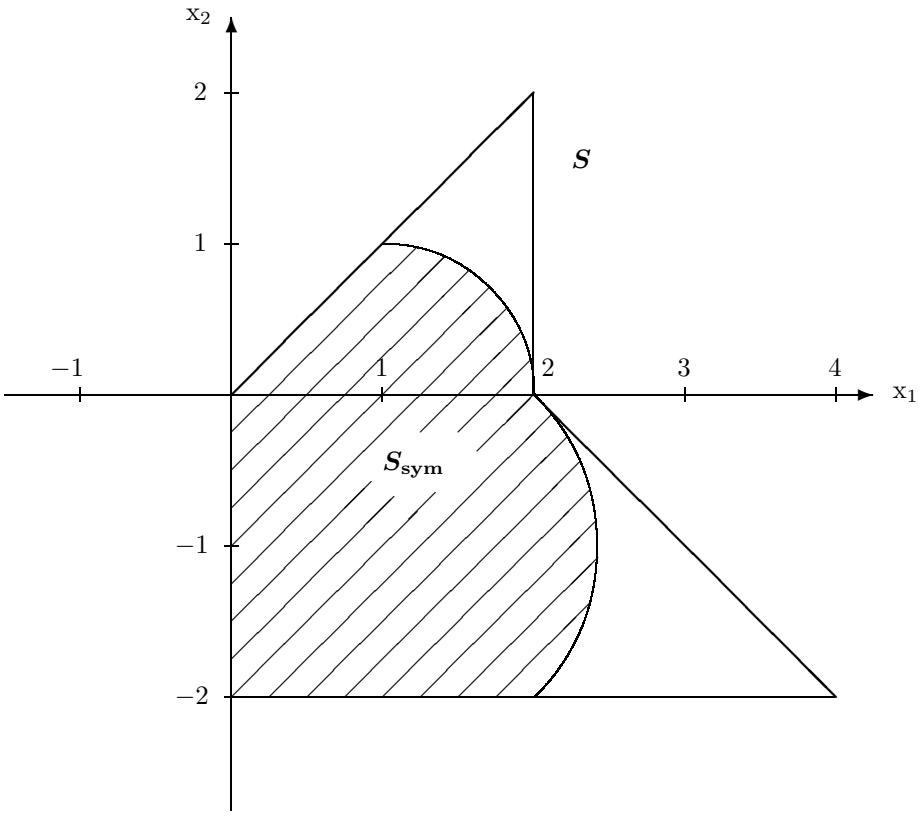


FIG. 4.2. *The shape of the solution sets S, S_{sym} in Example* 4.4.

Acknowledgment. We thank both referees for their valuable suggestions and remarks which improved this paper.

REFERENCES

[1] G. ALEFELD AND J. HERZBERGER, *Introduction to Interval Computations*, Academic Press, New York, 1983.

[2] G. ALEFELD, V. KREINOVICH, AND G. MAYER, *The shape of the symmetric solution set*, in Applications of Interval Computations, R. B. Kearfott and V. Kreinovich, eds., Kluwer, Boston, MA, 1995, pp. 61–79.

[3] G. ALEFELD, V. KREINOVICH, AND G. MAYER, *Symmetric linear systems with perturbed input data*, in Numerical Methods and Error Bounds, G. Alefeld and J. Herzberger, eds., Akademie Verlag, Berlin, 1996, pp. 16–22.

[4] G. ALEFELD AND G. MAYER, *On the symmetric and unsymmetric solution set of interval systems*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1223–1240.

[5] H. BEECK, *Über Struktur und Abschätzungen der Lösungsmenge von linearen Gleichungssystemen mit Intervallkoeffizienten*, Computing, 10 (1972), pp. 231–244.

[6] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[7] D. J. Hartfiel, *Concerning the solution set of $Ax = b$ where $P \leq A \leq Q$ and $p \leq b \leq q$*, Numer. Math., 35 (1980), pp. 355–359.
[8] C. Jansson, *Rigorous sensitivity analysis for real symmetric matrices with uncertain data*, in Computer Arithmetic, Scientific Computation and Mathematical Modelling, E. Kaucher, S. M. Markov, and G. Mayer, eds., Baltzer, Basel, Switzerland, 1991, pp. 293–316.
[9] C. Jansson, *Interval linear systems with symmetric matrices, skew-symmetric matrices and dependencies in the right hand side*, Computing, 46 (1991), pp. 265–274.
[10] A. Neumaier, *Interval Methods for Systems of Equations*, Cambridge University Press, Cambridge, UK, 1990.
[11] W. Oettli and W. Prager, *Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides*, Numer. Math., 6 (1964), pp. 405–409.
[12] J. Rohn, *Interval linear systems*, Freiburger Intervall–Berichte, 84/7 (1984), pp. 33–58.
[13] S. M. Rump, *Verification methods for dense and sparse systems of equations*, in Topics in Validated Computations, J. Herzberger, ed., Elsevier, Amsterdam, The Netherlands, 1994, pp. 63–135.
[14] A. Schrijver, *Theory of Linear and Integer Programming*, John Wiley, New York, 1986.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 3, pp. 706–732, July 1997

010

AN ANALYSIS OF SPECTRAL ENVELOPE REDUCTION VIA QUADRATIC ASSIGNMENT PROBLEMS*

ALAN GEORGE† AND ALEX POTHEN‡

Abstract. A new spectral algorithm for reordering a sparse symmetric matrix to reduce its envelope size was described in [Barnard, Pothen, and Simon, *Numer. Linear Algebra Appl.*, 2 (1995), pp. 317–334]. The ordering is computed by associating a Laplacian matrix with the given matrix and then sorting the components of a specified eigenvector of the Laplacian. In this paper we provide an analysis of the spectral envelope reduction algorithm. We describe related 1- and 2-sum problems; the former is related to the envelope size, while the latter is related to an upper bound on the work in an envelope Cholesky factorization. We formulate these two problems as quadratic assignment problems and then study the 2-sum problem in more detail. We obtain lower bounds on the 2-sum by considering a relaxation of the problem and then show that the spectral ordering finds a permutation matrix closest to an orthogonal matrix attaining the lower bound. This provides a stronger justification of the spectral envelope reduction algorithm than previously known. The lower bound on the 2-sum is seen to be tight for reasonably "uniform" finite element meshes. We show that problems with bounded separator sizes also have bounded envelope parameters.

Key words. 1-sum problem, 2-sum problem, envelope reduction, eigenvalues of graphs, Laplacian matrices, quadratic assignment problems, reordering algorithms, sparse matrices

AMS subject classifications. 65F50, 65K10, 68R10

PII. S089547989427470X

1. Introduction.

We provide a raison d'être for a novel spectral algorithm to reduce the envelope of a sparse symmetric matrix described in a companion paper [2]. The algorithm associates a discrete Laplacian matrix with the given symmetric matrix and then computes a reordering of the matrix by sorting the components of an eigenvector corresponding to the smallest nonzero Laplacian eigenvalue. The results in [2] show that the spectral algorithm can obtain significantly smaller envelope sizes compared to other currently used algorithms. All previous envelope reduction algorithms (known to us), such as the reverse Cuthill–McKee (RCM) algorithm and variants [3, 16, 17, 26, 37], are combinatorial in nature, employing breadth-first search to compute the ordering. In contrast, the spectral algorithm is an algebraic algorithm whose good envelope reduction properties are somewhat intriguing and poorly understood.

We describe problems related to envelope reduction called the 1- and 2-sum problems and then formulate these problems as quadratic assignment problems (QAPs). We show that the QAP formulation of the 2-sum enables us to obtain lower bounds

*Received by the editors September 26, 1994; accepted for publication (in revised form) by J. W. H. Liu July 28, 1996.

http://www.siam.org/journals/simax/18-3/27470.html

†Department of Computer Science, University of Waterloo, Waterloo, Ontario, N2L 3G1 Canada (jageorge@sparse1.uwaterloo.ca). The research of this author was supported by Canadian Natural Sciences and Engineering Research Council grant OGP0008111.

‡Department of Computer Science, Old Dominion University, Norfolk, VA 23529-0162 and ICASE, NASA Langley Research Center, Hampton, VA 23681-0001 (pothen@cs.odu.edu, pothen@icase.edu). The research of this author was supported by National Science Foundation grants CCR-9412698, DMS-9505110, and ECS-9527169, by U. S. Department of Energy grant DE-FG05-94ER25216, and by Canadian Natural Sciences and Engineering Research Council grant OGP0008111. This author was also supported by the National Aeronautics and Space Administration under NASA contract NAS1-19480 while he was in residence at the Institute for Computer Applications in Science and Engineering (ICASE).

on the 2-sum (and related envelope parameters) based on the Laplacian eigenvalues. The lower bounds seem to be quite tight for finite element problems when the mesh points are nearly all of the same degree and the geometries are simple. Further, a closest permutation matrix to an orthogonal matrix that attains the lower bound is obtained, to within a linear approximation, by sorting the second Laplacian eigenvector components in monotonically increasing or decreasing order. This justifies the spectral envelope-reducing algorithm more strongly than earlier results.

Although initially envelope-reducing orderings were developed for use in envelope schemes for sparse matrix factorization, these orderings have been used in the past few years in several other applications. The RCM ordering has been found to be an effective preordering in computing incomplete factorization preconditioners for preconditioned conjugate-gradient methods [4, 6]. Envelope-reducing orderings have been used in frontal methods for sparse matrix factorization [7].

The wider applicability of envelope-reducing orderings prompts us to take a fresh look at the reordering algorithms currently available and to develop new ordering algorithms. Spectral envelope reduction algorithms seem to be attractive in this context, since they

(i) compare favorably with existing algorithms in terms of the quality of the orderings [2],

(ii) extend easily to problems with weights, e.g., finite element meshes arising from discretizations of anisotropic problems, and

(iii) are fairly easily parallelizable.

Spectral algorithms are more expensive than the other algorithms currently available. But since the envelope reduction problem requires only one eigenvector computation (to low precision), we believe the costs are not impractically high in computation-intensive applications, e.g., frontal methods for factorization. In contexts where many problems having the same structure must be solved, a substantial investment in finding a good ordering might be justified since the cost can be amortized over many solutions. Improved algorithms that reduce the costs are being designed as well [25].

We focus primarily on the class of finite element meshes arising from discretizations of partial differential equations. Our goals in this project are to develop efficient software implementing our algorithms and to prove results about the quality of the orderings generated.

The projection approach for obtaining lower bounds of a QAP is due to Hadley, Rendl, and Wolkowicz [19], and this approach has been applied to the graph partitioning problem by the latter two authors [35]. In earlier work a spectral approach for the graph (matrix) partitioning problem has been employed to compute a spectral nested dissection ordering for sparse matrix factorization, for partitioning computations on finite element meshes on a distributed-memory multiprocessor [21, 33, 34, 36], and for load balancing parallel computations [22]. The spectral approach has also been used to find a pseudo peripheral node [18]. Juvan and Mohar [23, 24] have provided a theoretical study of the spectral algorithm for reducing p-sums, where $p = 1$, 2, and ∞, and Helmberg et al. [20] have obtained spectral lower bounds on the band width. A survey of some of these earlier results may be found in [31]. Paulino et al. [32] have also considered the use of spectral envelope reduction for finite element problems.

The following is an outline of the rest of this paper. In section 2 we describe various parameters of a matrix associated with its envelope and introduce the envelope size and envelope work minimization problems and the related 1- and 2-sum problems. We prove that bounds on the minimum 1-sum yield bounds on the minimum envelope

size and, similarly, bounds on the minimum 2-sum yield bounds on the work in an envelope Cholesky factorization. We also show in this section that minimizing the 2-sum is NP-complete. We compute lower bounds for the envelope parameters of a sparse symmetric matrix in terms of the eigenvalues of the Laplacian matrix in section 3. The popular RCM ordering is obtained by reversing the Cuthill–McKee (CM) ordering; the RCM ordering can never have a larger envelope size and work than the CM ordering, and is usually significantly better. We prove that reversing an ordering can improve or impair the envelope size by at most a factor Δ, and the envelope work by at most Δ^2, where Δ is the maximum degree of a vertex in the adjacency graph. In section 4, we formulate the 2- and 1-sum problems as QAPs. We obtain lower and upper bounds for the 2-sum problem in terms of the eigenvalues of the Laplacian matrix in section 5 by means of a projection approach that relaxes a permutation matrix to an orthogonal matrix with row and column sums equal to one. We justify the spectral envelope reduction algorithm in section 6 by proving that a closest permutation matrix to an orthogonal matrix attaining the lower bound for the 2-sum is obtained, to within a linear approximation of the problem, by permuting the second Laplacian eigenvector in monotonically increasing or decreasing order. In section 7 we show that graphs with small separators have small envelope parameters as well by considering a modified nested dissection ordering. We present computational results in section 8 to illustrate that the 2-sums obtained by the spectral reordering algorithm can be close to optimal for many finite element meshes. Section 9 contains our concluding remarks. The appendix contains some lower bounds for the more general p-sum problem, where $1 \leq p < \infty$.

2. A menagerie of envelope problems.

2.1. The envelope of a matrix. Let A be an $n \times n$ symmetric matrix with elements a_{ij}, whose diagonal elements are nonzero. Various parameters of the matrix A associated with its envelope are defined below.

We denote the column indices of the nonzeros in the lower triangular part of the ith row by

$$\text{row}(i) = \{j : a_{ij} \neq 0 \text{ and } 1 \leq j \leq i\}.$$

For the ith row of A we define

$$f_i(A) = \min\{j : j \in \text{row}(i)\} \quad \text{and}$$
$$r_i(A) = i - f_i(A).$$

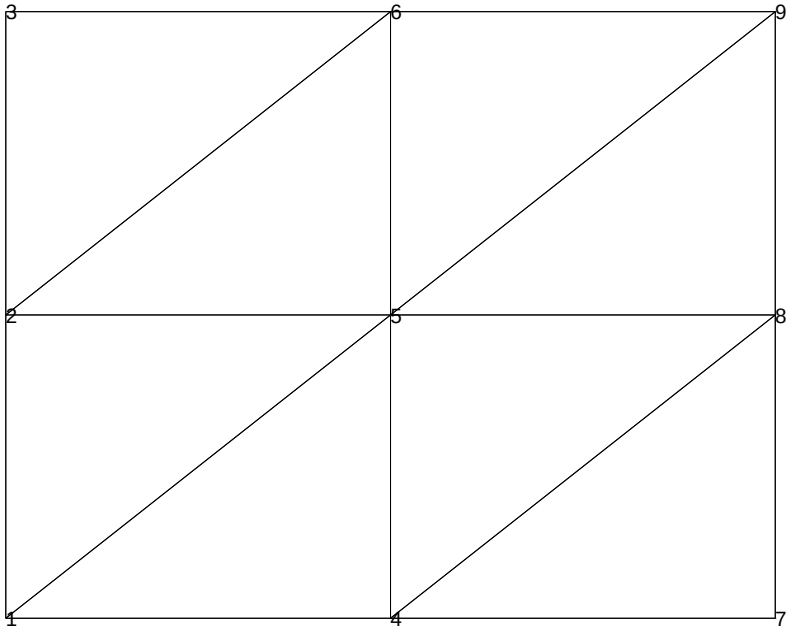
Here $f_i(A)$ is the column index of the first nonzero in the ith row of A (by our assumption of nonzero diagonals, $1 \leq f_i \leq i$) and the parameter $r_i(A)$ is the *row width* of the ith row of A. The *bandwidth* of A is the maximum row width

$$\text{bw}(A) = \max\{r_i(A) : i = 1, \ldots, n\}.$$

The *envelope* of A is the set of index pairs

$$\text{Env}(A) = \{(i, j) : f_i(A) \leq j < i,\ i = 1, \ldots, n\}.$$

For each row, the column indices lie in an interval beginning with the column index of the first nonzero element and ending with (but not including) the index of the diagonal nonzero element.



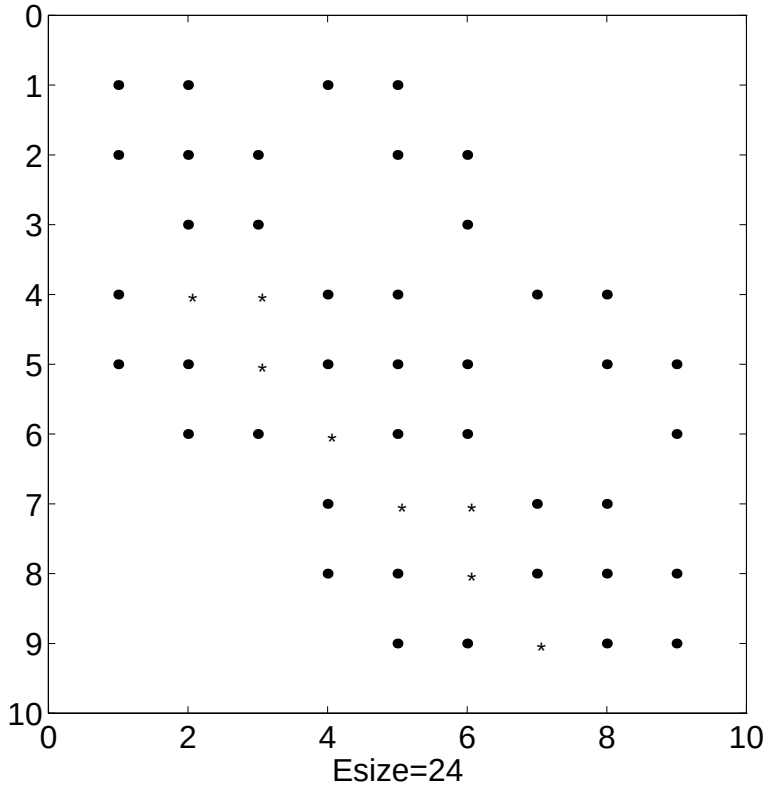


FIG. 2.1. *An ordering of a 7-point grid and the corresponding matrix. The lower triangle of the envelope is indicated by marking zeros within it by asterisks.*

TABLE 2.1
Row widths and column widths of the matrix in Figure 2.1.

i	1	2	3	4	5	6	7	8	9
r_i	0	1	1	3	4	4	3	4	4
c_i	3	4	3	4	4	3	2	1	0

We denote the size of the envelope by $\text{Esize}(A) = |\text{Env}(A)|$. (The number $\text{Esize}(A) + n$ (which includes the diagonal elements) is called the *profile* of A [7].) The work in the Cholesky factorization of A that employs an envelope storage scheme is bounded from above by

$$\text{Wbound}(A) \equiv (1/2) \sum_{i=2}^{n} r_i(r_i + 3).$$

This bound is tight [29] when an ordering satisfies (1) $f_i(A) \leq f_j(A)$ when $i < j$ for all i, j between 1 and n, and (2) $f_i(A) < i$ for all $i = 2, \ldots, n$.

A 3×3 7-point grid and the nonzero structure of the corresponding matrix A are shown in Figure 2.1. A "$\bullet$" indicates a nonzero element and a "$*$" indicates a zero element that belongs to the lower triangle of the envelope in the matrix. The row widths given in Table 2.1 are easily verified from the structure of the matrix. The envelope size is obtained by summing the row widths and is equal to 24. (Column widths c_i are defined later in this section.)

The values of these parameters strongly depend on the choice of an ordering of the rows and columns. Hence we consider how these parameters vary over symmetric permutations $P^T AP$ of a matrix A, where P is a permutation matrix. We define $\text{Esize}_{\min}(A)$, the minimum envelope size of A, to be the minimum envelope size among all permutations $P^T AP$ of A. The quantities $\text{Wbound}_{\min}(A)$ and $\text{bw}_{\min}(A)$ are defined in a similar fashion. Minimizing the envelope size and the bandwidth of a matrix are NP-complete problems [28], and minimizing the work bound is likely to be intractable as well. So one must settle for heuristic orderings to reduce these quantities.

It will be helpful in section 3 to consider a "column-oriented" expression for the envelope size for obtaining a lower bound on this quantity. The *width* of a column j of A is the number of row indices in the jth column of the envelope of A. In other words,

$$c_j(A) = |\{k : k > j \text{ and } \exists \ell \leq j \ni a_{k\ell} \neq 0\}|.$$

(This is also called the jth *front width.*) It is then easily seen that the envelope size is

$$\text{Esize}(A) = \sum_{j=1}^{n} c_j. \tag{2.1}$$

The work in an envelope factorization scheme is given by

$$\text{Ework}(A) = (1/2) \sum_{j=1}^{n} c_j^2, \tag{2.2}$$

where we have ignored the linear term in c_j. The column widths of the matrix in Figure 2.1 are given in Table 2.1. These concepts and their interrelationships are described by Liu and Sherman [29] and are also discussed in [5, 15].

The envelope parameters can also be defined with respect to the adjacency graph $G = (V, E)$ of A. Denote $\mathrm{nbr}(v) = \{v\} \cup \mathrm{adj}(v)$. In terms of the graph G and an ordering α of its vertices, we can define

$$r(v, \alpha) = \max\{\alpha(v) - \alpha(w) : w \in \mathrm{nbr}(v), \alpha(w) \leq \alpha(v)\}.$$

Hence we can write the envelope size and work associated with an ordering α as

$$\mathrm{Esize}(G, \alpha) = \sum_{v \in V} r(v, \alpha) \;=\; \sum_{v \in V} \max\{\alpha(v) - \alpha(w) : w \in \mathrm{nbr}(v), \alpha(w) \leq \alpha(v)\},$$

$$\mathrm{Wbound}(G, \alpha) = \sum_{v \in V} r(v, \alpha)^2 \;=\; \sum_{v \in V} \max\{(\alpha(v) - \alpha(w))^2 : w \in \mathrm{nbr}(v), \alpha(w) \leq \alpha(v)\}.$$

The goal is to choose a vertex ordering $\alpha : V \mapsto \{1, \ldots, n\}$ to minimize one of the parameters described above. We denote by $\mathrm{Esize}_{\min}(G)$ ($\mathrm{Wbound}_{\min}(G)$) the minimum value of $\mathrm{Esize}(G, \alpha)$ ($\mathrm{Wbound}(G, \alpha)$) over all orderings α. The reader can compute the envelope size of the numbered graph in Figure 2.1 using the definition given in this paragraph, to verify that $\mathrm{Esize}(G) = 24$.

The jth front width has an especially nice interpretation if we consider the adjacency graph $G = (V, E)$ of A. Let the vertex corresponding to a column j of A be numbered v_j, so that $V = \{v_1, \ldots, v_n\}$, and define $V_j = \{v_1, \ldots, v_j\}$. Denote $\mathrm{adj}(X) = (\cup_{v \in X} \mathrm{adj}(v)) \setminus X$ for a subset of vertices X. Then $c_j(A) = |\mathrm{adj}(V_j)|$.

To illustrate the dependence of the envelope size on the ordering, we include in Figure 2.2 an ordering that leads to a smaller envelope size for the 7-point grid. Again, a "$\bullet$" indicates a nonzero element and a "$*$" indicates a zero element that belongs to the lower triangle of the envelope in the matrix. This ordering by "diagonals" yields the optimal envelope size for the 7-point grid [27].

2.2. 1- and 2-sum problems. It will be helpful to consider quantities related to the envelope size and envelope work, the 1-sum and the 2-sum.

For real $1 \leq p < \infty$, we define the p-sum to be

$$\sigma_p^p(A) = \sum_{i=1}^{n} \sum_{j \in \mathrm{row}(i)} (i - j)^p.$$

Minimizing the 1-sum ($p = 1$) is the *optimal linear arrangement problem*, and the limiting case $p = \infty$ corresponds to the minimum *bandwidth problem*; both these are well-known NP-complete problems [13]. We will show in the section 2.3 that minimizing the 2-sum is NP-complete as well.

We write the envelope size and the 1-sum, and the envelope work and the 2-sum, in a way that shows their relationships:

$$\mathrm{Esize}(A) = \sum_{i=1}^{n} \max_{j \in \mathrm{row}(i)} (i - j), \tag{2.3}$$

$$\sigma_1(A) = \sum_{i=1}^{n} \sum_{j \in \mathrm{row}(i)} (i - j); \tag{2.4}$$

$$\mathrm{Wbound}(A) = \sum_{i=1}^{n} \max_{j \in \mathrm{row}(i)} (i - j)^2, \tag{2.5}$$

$$\sigma_2^2(A) = \sum_{i=1}^{n} \sum_{j \in \mathrm{row}(i)} (i - j)^2. \tag{2.6}$$

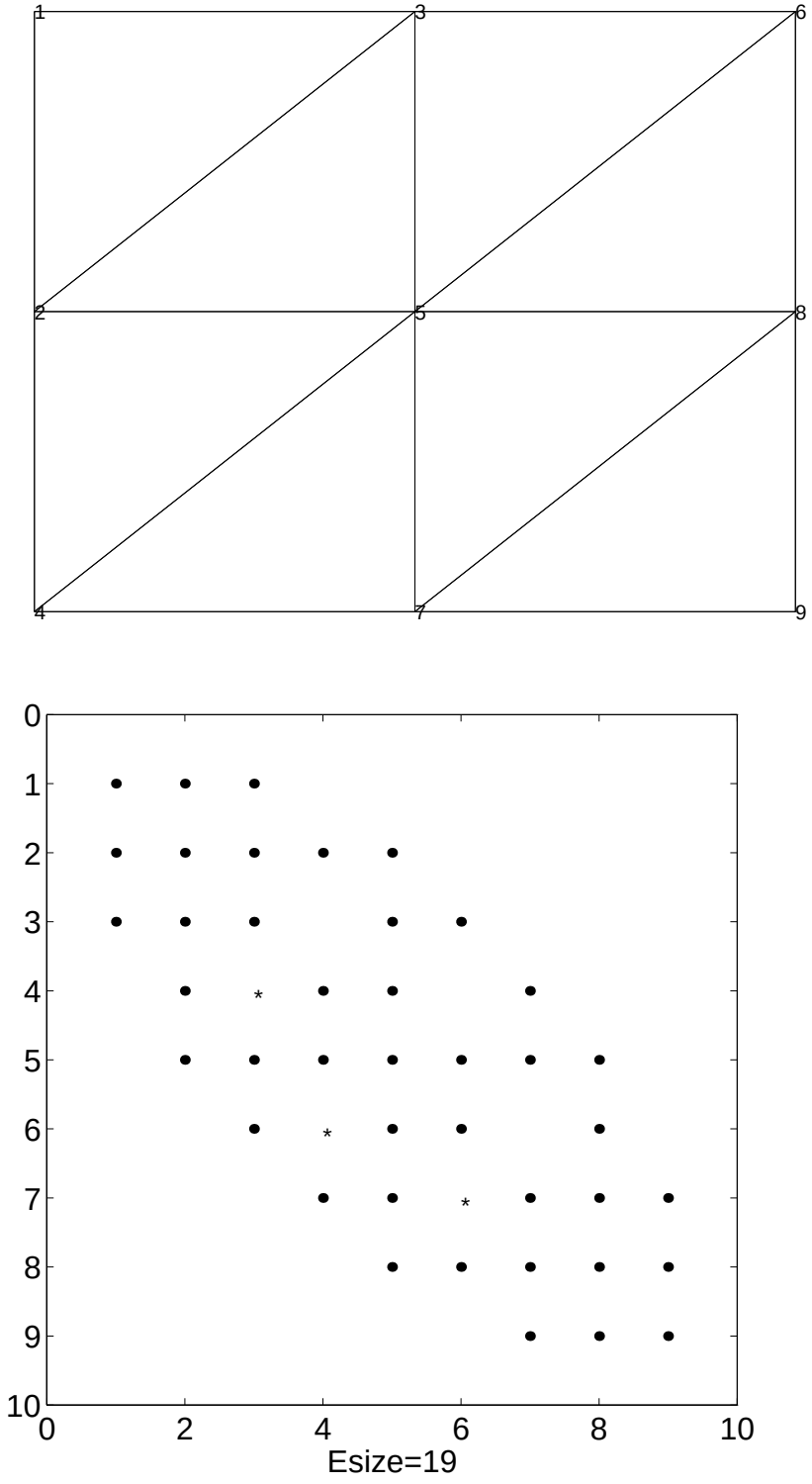


FIG. 2.2. *Another ordering of a 7-point grid and the corresponding matrix. Again the lower triangle of the envelope is indicated by marking the zeros within it by asterisks.*

The parameters $\sigma_{1,\min}(A)$ and $\sigma^2_{2,\min}(A)$ are the minimum values of these parameters over all symmetric permutations P^TAP of A.

We now consider the relationships between bounds on the envelope size and the 1-sum and between the upper bound on the envelope work and the 2-sum. Let Δ denote the maximum number of off-diagonal nonzeros in a row of A. (This is the maximum vertex degree in the adjacency graph of A.)

THEOREM 2.1. *The minimum values of the envelope size, envelope work in the Cholesky factorization, 1-sum, and 2-sum of a symmetric matrix A are related by the following inequalities:*

$$\text{Esize}_{\min}(A) \le \sigma_{1,\min}(A) \le \Delta\text{Esize}_{\min}(A), \tag{2.7}$$

$$\text{Wbound}_{\min}(A) \le \sigma^2_{2,\min}(A) \le \Delta\text{Wbound}_{\min}(A), \tag{2.8}$$

$$\sigma_{2,\min}(A) \le \sigma_{1,\min}(A) \le \sqrt{|E|}\,\sigma_{2,\min}(A). \tag{2.9}$$

Proof. We begin by proving (2.8). Our strategy will be first to prove the inequalities

$$\text{Wbound}(A) \le \sigma^2_2(A) \le \Delta\text{Wbound}(A),$$

and then to obtain the required result by considering two different permutations of A.

The bound $\text{Wbound}(A) \le \sigma^2_2(A)$ is immediate from equations (2.5) and (2.6). If the inner sum in the latter equation is bounded from above by

$$\Delta \max_{j\in row(i)} (i-j)^2,$$

then we get $\Delta\text{Wbound}(A)$ as an upper bound on the 2-sum.

Now let X_1 be a permutation matrix such that $\widetilde{A_1} \equiv X_1^TAX_1$ and $\text{Wbound}(\widetilde{A_1}) = \text{Wbound}_{\min}(A)$. Then we have

$$\sigma^2_{2,\min}(A) \le \sigma^2_2(\widetilde{A_1}) \le \Delta\text{Wbound}(\widetilde{A_1}) = \Delta\text{Wbound}_{\min}(A).$$

Further, let X_2 be a permutation matrix such that $\widetilde{A_2} \equiv X_2^TAX_2$ and $\sigma^2_2(\widetilde{A_2}) = \sigma^2_{2,\min}(A)$. Again, we have

$$\text{Wbound}_{\min}(A) \le \text{Wbound}(\widetilde{A_2}) \le \sigma^2_2(\widetilde{A_2}) = \sigma^2_{2,\min}(A).$$

We obtain the result by putting the last two inequalities together.

We omit the proof of (2.7) since it can be obtained by a similar argument and proceed to prove (2.9). The first inequality $\sigma_2(A) \le \sigma_1(A)$ holds since the p-norm of any real vector is a decreasing function of p. The second inequality is also standard since it bounds the 1-norm of a vector by means of its 2-norm. This result was obtained earlier by Juvan and Mohar [24]; we include its proof for completeness. Applying the Cauchy–Schwarz inequality to $\sigma^2_1(A)$, we have

$$\left(\sum_{i=1}^{n}\sum_{j\in\text{row}(i)}(i-j)\right)^2$$
$$\le \left(\sum_{i=1}^{n}\sum_{j\in\text{row}(i)}1\right)\left(\sum_{i=1}^{n}\sum_{j\in\text{row}(i)}(i-j)^2\right) = |E|\sigma^2_2(A).$$

We obtain the result by considering two orderings that achieve the minimum 1- and 2-sums. □

2.3. Complexity of the 2-sum problem. We proceed to show that minimizing the 2-sum is NP-complete. In section 8 we show that the spectral algorithm computes a 2-sum within a factor of two for the finite element problems in our test collection. This proof shows that despite the near-optimal solutions obtained by the spectral algorithm on this test set, it is unlikely that a polynomial time algorithm can be designed for computing the minimum 2-sum.

Readers who are willing to accept the complexity of this problem without proof should skip this section; we recommend that everyone do so on a first reading.

Given a graph $G = (V, E)$ on n vertices, MINTWOSUM is the problem of deciding if there exists a numbering of its vertices $\alpha : V \mapsto \{1, \ldots, n\}$ such that $\sum_{(u,v)\in E}(\alpha(u)-\alpha(v))^2 \leq k$ for a given positive integer k. This is the decision version of the problem of minimizing the 2-sum of G.

THEOREM 2.2. *MINTWOSUM is NP-complete.*

Remark. This proof follows the framework for the NP-completeness of the 1-sum problem in Even [8, section 10.7], but the details are substantially different.

Proof. The theorem will follow if we show that MAXTWOSUM, the problem of deciding whether a graph G' on n vertices has a vertex numbering with 2-sum *greater than or equal to* a given positive integer k', is NP-complete. For, the 2-sum of G' under some ordering is at least k' if and only if the 2-sum of the complement of G' under the same ordering is at most $p(n) - k'$, where $p(n) = \sum_{j=1}^{n}\sum_{i=1}^{j-1}(j-i)^2 = n^4/12 - n^2/12$ is the 2-sum of the complete graph.

We show that MAXTWOSUM is NP-complete by a reduction from MAXCUT, the problem of deciding whether a given graph $G = (V, E)$ has a partition of its vertices into two sets $\{S, V \setminus S\}$ such that $|\delta(S, V \setminus S)|$, the number of edges joining S and $V \setminus S$, is at least a given positive integer k. From the graph G we construct a graph $G' = (V' \equiv V \cup \{x_1, \ldots, x_{n^4}\}, E' \equiv E)$ by adding n^4 isolated vertices to V and no edges to E. We claim that G has a cut of size at least k if and only if G' has a 2-sum at least $k' \equiv k \cdot n^8$.

If G has a cut $(S, V \setminus S)$ of size at least k, define an ordering α' of G' by interposing the n^4 isolated vertices between S and $V \setminus S$: number the vertices in S first, the isolated vertices next, and the vertices in $V \setminus S$ last, where the ordering among the vertices in each set S and $V \setminus S$ is arbitrary. Every edge belonging to the cut contributes at least n^8 to the 2-sum, and hence its value is at least $k \cdot n^8$.

The converse is a little more involved.

Suppose that G' has an ordering $\alpha' : V' \mapsto \{1, 2, \ldots, n + n^4\}$ with 2-sum greater than or equal to $k \cdot n^8$. The ordering α' of G' induces a natural ordering $\alpha : V \mapsto \{1, \ldots, n\}$ of G if we ignore the isolated vertices and maintain the relative ordering of the vertices in V. For each $1 \leq i \leq n$, define the ordered set $S_i = \{v \in V : \alpha(v) \leq i\}$. Then each pair $(S_i, V \setminus S_i)$ is a cut in G. Further, each such cut in G induces a cut $(S_i', V' \setminus S_i')$ in the larger graph G' as follows. The vertex set S_i' is formed by augmenting S_i with the isolated vertices numbered lower than the highest numbered (nonisolated) vertex in S_i (with respect to the ordering α').

We now choose a cut $(S', V' \setminus S')$ that maximizes the "1-sum over the cut edges"

$$\sum_{\substack{v\in S', w\in V'\setminus S' \\ (v,w)\in E',}} |\alpha'(v) - \alpha'(w)|,$$

from among the n cuts $(S'_i, V' \setminus S'_i)$. By means of this cut and the ordering α', we define a new ordering β' by moving the isolated vertices in the ordered set S' to the highest numbers in that set, by moving the isolated vertices in $V' \setminus S'$ to the lowest numbers in that set, and by preserving the relative ordering of the other vertices. The effect is to interpose the isolated vertices in "between" the two sets of the cut.

Claim. The 2-sum of the graph G' under the ordering β' is greater than that under α'.

To prove the claim, we examine what happens when an isolated vertex x belonging to S' is moved to the higher end of that ordered set.

Define three sets A', B', C' as follows. The set A' (B') is the set of vertices in S' numbered lower (higher) than x in the ordering α' and $C' \equiv V' \setminus S'$. Also, let E_1 denote the edges joining A' and B', E_2 denote edges joining B' and C', and E_3 denote those joining A' and C'.

Denote the contribution, with respect to the ordering α', of an edge $e_k \in E_1$ to the 1-sum by a_k and that of an edge $e_l \in E_2$ by b_l. Then the change in the 2-sum due to moving x is

$$\begin{aligned}&\sum_{E_2}(b_l+1)^2 - b_l^2 \;+\; \sum_{E_1}(a_k-1)^2 - a_k^2\\ &= |E_1| + |E_2| + \sum_{E_2} 2b_l - \sum_{E_1} 2a_k.\end{aligned}$$

The third term on the right-hand side is the contribution to the 1-sum made by the edges E_2 in the cut $(A' \cup B', C') \equiv (S', V' \setminus S')$, while the fourth term is the contribution made by the edges E_1 in the cut $(A', B' \cup C')$. By the choice of the cut $(S', V' \setminus S')$, we find that the difference is positive, and hence that the 2-sum has increased in the new ordering obtained from α' by moving the vertex x.

We now show that after moving the vertex x, $(A' \cup B', C')$ continues to be a cut that maximizes the 1-sum over the cut edges among all cuts $(S'_i, V' \setminus S'_i)$ *with respect to the new ordering.* For this cut, the 1-sum over cut edges has increased by $|E_2|$ because the number of each vertex in B has decreased by one in the new ordering. Among cuts with one set equal to an ordered subset of A', the 1-sum over cut edges can only decrease when x is moved, since the set B' moves closer to A', and C' does not move at all relative to A'. Now consider cuts of the form $(A' \cup B'_1, B'_2 \cup C')$, with B'_1 an ordered subset of B', and $B'_1 \cup B'_2 = B'$. The cut edges now join A' to $B'_2 \cup C'$, and B'_1 to $B'_2 \cup C'$. The edges joining A' to B'_2 contribute a smaller value to the 1-sum in the new ordering relative to α', while the edges joining A' to C' contribute the same to the 1-sum in both cuts $(A' \cup B', C')$ and $(A' \cup B'_1, B'_2 \cup C')$ under the new ordering. The edges joining B'_1 and B'_2 do not change their contribution to the 1-sum in the new ordering. The edges that join B'_1 and C' form a subset of the edges that join B' and C', and hence the contribution of the former to the 1-sum is no larger than the contribution of the latter set in the new ordering. This shows that the cut $(A' \cup B', C')$ continues to have a 1-sum over the cut edges larger than or equal to that of any cut $(A' \cup B'_1, B'_2 \cup C')$. Finally, any cut that includes A', B', and an ordered subset C'_1 of C' can be shown by similar reasoning to not have a larger 1-sum than $(S', V' \setminus S')$.

The reasoning in the previous paragraph permits us to move the isolated vertices in S' one by one to the higher end of that set without decreasing the 2-sum while simultaneously preserving the condition that the cut $(S', V' \setminus S')$ has the maximum value of the 1-sum over the cut edges. The argument that we can move the isolated

vertices in $V' \setminus S'$ to the beginning of that ordered set follows from symmetry since both the 2-sum and the 1-sum are unchanged when we reverse an ordering. Hence by inducting over the number of isolated vertices moved, the ordering β' has a 2-sum at least as large as the ordering α'. This completes the proof of the claim.

The rest of the proof involves computing an upper bound on the 2-sum of the graph G' under the ordering β' to show that since G' has 2-sum greater than k', the graph G has a cut of size at least k.

Let $\delta \equiv |(S', V' \setminus S')|$. The cut edges contribute at most $\delta \cdot (n^4 + n)^2$ to the upper bound on the 2-sum; the uncut edges contribute at most the 2-sum of a complete graph on n vertices. The latter is $p(n) \equiv n^4/12 - n^2/12$. Thus we have, keeping only the positive terms,

$$\begin{aligned} &\delta(n^4 + n)^2 + (1/12)n^4 \geq kn^8 \\ \Rightarrow\ &\delta + (2\delta)/n^3 + (1/(12n^4)) + \delta/n^6 \geq k. \end{aligned}$$

Since the number of cut edges δ is at most $n^2/2$, the sum of the latter three terms on the left-hand side is easily computed to be less than 1 for $n > 2$. Hence we conclude that the graph G has a cut with at least k edges. This completes the proof of the theorem. □

3. Bounds for envelope size. In this section we present lower bounds for the minimum envelope size and the minimum work involved in an envelope Cholesky factorization in terms of the second Laplacian eigenvalue. We will require some background on the Laplacian matrix.

3.1. The Laplacian matrix. The Laplacian matrix $Q(G)$ of a graph G is the $n \times n$ matrix $D - M$, where D is the diagonal degree matrix and M is the adjacency matrix of G. If G is the adjacency graph of a symmetric matrix A, then we could define the Laplacian matrix Q directly from A:

$$q_{ij} = \begin{cases} -1 & \text{if } i \neq j \text{ and } a_{ij} \neq 0, \\ 0 & \text{if } i \neq j \text{ and } a_{ij} = 0, \\ \sum_{\substack{j=1 \\ j \neq i}}^{n} |q_{ij}| & \text{if } i = j. \end{cases}$$

Note that

$$\begin{aligned} \underline{x}^T Q \underline{x} &= \underline{x}^T D \underline{x} - \underline{x}^T M \underline{x} \\ &= \sum_{\substack{j \leq i \\ a_{ij} \neq 0}} (x_i - x_j)^2. \end{aligned} \tag{3.1}$$

The eigenvalues of $Q(G)$ are the *Laplacian eigenvalues* of G, and we list them as $\lambda_1(Q) \leq \lambda_2(Q) \leq \cdots \leq \lambda_n(Q)$. An eigenvector corresponding to $\lambda_k(Q)$ will be denoted by $\underline{x}_k$ and will be called a kth eigenvector of Q. It is well known that Q is a singular M-matrix, and hence its eigenvalues are nonnegative. Thus $\lambda_1(Q) = 0$, and the corresponding eigenvector is any nonzero constant vector $\underline{c}$. If G is connected, then Q is irreducible, and then $\lambda_2(Q) > 0$; the smallest nonzero eigenvalues and the corresponding eigenvectors have important properties that make them useful in the solution of various partitioning and ordering problems. These properties were first investigated by Fiedler [9, 10]; as discussed in section 1, more recently several authors have studied their application to such problems.

3.2. Laplacian bounds for envelope parameters. It will be helpful to work with the "column-oriented" definition of the envelope size. Let the vertex corresponding to a column j of A be numbered v_j in the adjacency graph so that $V = \{v_1, \ldots, v_n\}$, and let $V_j = \{v_1, \ldots, v_j\}$. Recall that the *column width* of a vertex v_j is $c_j = |\mathrm{adj}(V_j)|$ and that the envelope size of G (or A) is

$$\mathrm{Esize}(G) = \sum_{j=1}^{n} c_j.$$

Recall also that Δ denotes the maximum degree of a vertex. Given a set of vertices S, we denote by $\delta(S)$ the set of edges with one endpoint in S and the other in $V \setminus S$.

We make use of the following elementary result, where the lower bound is due to Alon and Milman [1] and the upper bound is due to Juvan and Mohar [24].

LEMMA 3.1. *Let $S \subset V$ be a subset of the vertices of a graph G. Then*

$$\lambda_2(Q)\frac{|S||V \setminus S|}{n} \le |\delta(S)| \le \lambda_n(Q)\frac{|S||V \setminus S|}{n}. \qquad \square$$

THEOREM 3.2. *The envelope size of a symmetric matrix A can be bounded in terms of the eigenvalues of the associated Laplacian matrix as*

$$\frac{\lambda_2(Q)}{6\Delta}(n^2 - 1) \le \mathrm{Esize}(A) \le \frac{\lambda_n(Q)}{6}(n^2 - 1).$$

Proof. From Lemma 3.1,

$$|\delta(V_j)| \ge \frac{\lambda_2(Q)}{n} j(n - j).$$

Now $c_j(A) = |\mathrm{adj}(V_j)| \ge |\delta(V_j)|/\Delta$; substituting the lower bound for $|\delta(V_j)|$ and summing this latter expression over all j, we obtain the lower bound on the envelope size.

The upper bound is obtained by using the inequality $c_j(A) \le |\delta(V_j)|$ with the upper bound in Lemma 3.1. $\square$

A lower bound on the work in an envelope Cholesky factorization can be obtained from the lower bound on the envelope size.

THEOREM 3.3. *A lower bound on the work in the envelope Cholesky factorization of a symmetric positive-definite matrix A is*

$$\mathrm{Ework}(A) \ge \frac{\mathrm{Esize}(A)^2}{2n}.$$

Proof. The proof follows from equations (2.1) and (2.2) by an application of the Cauchy–Schwarz inequality. We omit the details. $\square$

Cuthill and McKee [3] proposed one of the earliest ordering algorithms for reducing the envelope size of a sparse matrix. George [14] discovered that reversing this ordering leads to a significant reduction in envelope size and work. The envelope parameters obtained from the RCM ordering are never larger than those obtained from CM [29]. The RCM ordering has become one of the most popular envelope size reducing orderings. However, we do not know of any published quantitative results on the improvement that may be expected by reversing an ordering, and here we present

the first such result. For degree-bounded finite element meshes, no asymptotic improvement is possible; the parameters are improved only by a constant factor. Of course, in practice, a reduction by a constant factor could be quite significant.

THEOREM 3.4. *Reversing the ordering of a sparse symmetric matrix A can change (improve or impair) the envelope size by at most a factor Δ and the envelope work by at most Δ^2.*

Proof. Let v_j denote the vertex in the adjacency graph corresponding to the jth column of A (in the original ordering) so that the jth column width $c_j(A) = |\mathrm{adj}(V_j)|$, where $V_j = \{v_1, \ldots, v_j\}$. Let $\widetilde{A}$ denote the permuted matrix obtained by reversing the column and row ordering of A. We have the inequality

$$c_j(A) = |\mathrm{adj}(V_j)| \leq |\delta(V_j)| \leq \Delta |\mathrm{adj}(V \setminus V_j)| = \Delta c_{n-j}(\widetilde{A}).$$

Since $\mathrm{Esize}(A) = \sum_{j=1}^{n} c_j(A)$, summing this inequality over j from one to n, we obtain $\mathrm{Esize}(A) \leq \Delta \mathrm{Esize}(\widetilde{A})$. By symmetry, the inequality $\mathrm{Esize}(\widetilde{A}) \leq \Delta \mathrm{Esize}(A)$ holds as well.

The inequality on the envelope work follows by a similar argument from the equation $\mathrm{Ework}(A) = (1/2) \sum_{j=1}^{n} c_j^2$. □

4. Quadratic assignment formulation of 2- and 1-sum problems. We formulate the 2- and 1-sum problems as QAPs in this section.

4.1. The 2-sum problem. Let the vector $\underline{p} = \begin{pmatrix} 1 & 2 & \cdots & n \end{pmatrix}^T$, and let $\underline{\alpha}$ be a permutation vector, i.e., a vector whose components form a permutation of 1, …, n. We may write $\underline{\alpha} = X\underline{p}$, where X is a permutation matrix with elements

$$x_{ij} = \begin{cases} 1 & \text{if } j = \alpha(i), \\ 0 & \text{otherwise.} \end{cases}$$

It is easily verified that the $(\alpha(i), \alpha(j))$ element of the permuted matrix $X^T A X$ is the element a_{ij} of the unpermuted matrix A. Let $B = \underline{p}\,\underline{p}^T$; then $b_{ij} = ij$. We denote the set of all permutation vectors with n components by S_n.

We write the 2-sum as a quadratic form involving the Laplacian matrix Q:

$$\begin{aligned} \sigma^2_{2,\min}(A) &= \min_X \sigma_2^2(X^T A X) \\ &= \min_{\underline{\alpha} \in S_n} \sum_{\substack{\alpha(j) \leq \alpha(i) \\ a_{\alpha(i),\alpha(j)} \neq 0}} (\alpha(i) - \alpha(j))^2 \\ &= \min_{\underline{\alpha} \in S_n} \underline{\alpha}^T Q \underline{\alpha} \\ &= \min_{\underline{\alpha} \in S_n} \sum_{i=1}^{n} \sum_{j=1}^{n} q_{ij}\, \alpha(i)\alpha(j). \end{aligned}$$

The transformation from the second to the third line makes use of (3.1).

This quadratic form can be expressed as a QAP by substituting $b_{\alpha(i),\alpha(j)} = \alpha(i)\alpha(j)$:

$$\min_{\underline{\alpha} \in S_n} \underline{\alpha}^T Q \underline{\alpha} = \min_{\underline{\alpha} \in S_n} \sum_{i=1}^{n} \sum_{j=1}^{n} q_{ij}\, b_{\alpha(i)\alpha(j)}.$$

There is also a trace formulation of the QAP in which the variables are the elements of the permutation matrix X. We obtain this formulation by substituting $X\underline{p}$ for $\underline{\alpha}$. Thus

$$\min_{\underline{\alpha}\in S_n} \underline{\alpha}^T Q \underline{\alpha} = \min_X \underline{p}^T X^T Q X \underline{p}.$$

We may consider the last scalar expression as the trace of a 1×1 matrix and then use the identity $\operatorname{tr} MN = \operatorname{tr} NM$ to rewrite the right-hand side of the last displayed equation as

$$\min_X \operatorname{tr} QX\underline{p}\,\underline{p}^T X^T \equiv \min_X \operatorname{tr} QXBX^T.$$

This is a QAP since it is a quadratic in the unknowns x_{ij}, which are the elements of the permutation matrix X. The fact that B is a rank-one matrix leads to great simplifications and savings in the computation of good lower bounds for the 2-sum problem.

4.2. The 1-sum problem. Let M be the adjacency matrix of a given symmetric matrix A and let S denote a "distance matrix" with elements $s_{ij} = |i - j|$, both of order n. Then

$$\begin{aligned}
\sigma_{1,\min}(A) &= \min_X \sigma_1(X^T A X) \\
&= \min_{\underline{\alpha}\in S_n} \sum_{\substack{\alpha(j)\le\alpha(i) \\ m_{\alpha(i),\alpha(j)}\neq 0}} \alpha(i) - \alpha(j) \\
&= (1/2) \min_{\underline{\alpha}\in S_n} \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}\, s_{\alpha(i),\alpha(j)} \\
&= (1/2) \min \operatorname{tr} MXSX^T.
\end{aligned}$$

Unlike the 2-sum, the matrices involved in the QAP formulation of the 1-sum are both of rank n. Hence the bounds we obtain for this problem by this approach are considerably more involved, and will not be considered here.

5. Eigenvalue bounds for the 2-sum problem.

5.1. Orthogonal bounds. A technique for obtaining lower (upper) bounds for the QAP

$$\min_X \operatorname{tr} QXBX^T, \quad X \text{ is a permutation matrix},$$

is to relax the requirement that the minimum (maximum) be attained over the class of permutation matrices. Let $\underline{u} = (1/\sqrt{n})\begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix}$ denote the normalized n-vector of all ones. A matrix X of order n is a permutation matrix if and only if it satisfies the following three constraints:

$$X\underline{u} = \underline{u}, \quad X^T\underline{u} = \underline{u}, \tag{5.1}$$

$$X^T X = I_n, \tag{5.2}$$

$$x_{ij} \ge 0, \ i, j = 1, \dots, n. \tag{5.3}$$

The first of these, the *stochasticity constraint*, expresses the fact that each row sum or column sum of a permutation matrix is one; the second states that a permutation

matrix is orthogonal; the third states that its elements are nonnegative. The simplest bounds for a QAP are obtained when we relax both the stochasticity and nonnegativity constraints and insist only that X be orthonormal. The following result is from [11]; see also [12].

THEOREM 5.1. *Let the eigenvalues of a matrix be ordered*

$$\lambda_1(\cdot) \le \lambda_2(\cdot) \le \cdots \le \lambda_n(\cdot).$$

Then, as X varies over the set of orthogonal matrices, the following upper and lower bounds hold:

$$\sum_{i=1}^{n} \lambda_i(Q)\lambda_{n+1-i}(B) \le \operatorname{tr} QXBX^T \le \sum_{i=1}^{n} \lambda_i(Q)\lambda_i(B). \qquad \square$$

The Laplacian matrix Q has $\lambda_1(Q) = 0$; also $\lambda_i(B) = 0$, for $i = 1, \ldots, n-1$, and $\lambda_n(B) = \underline{p}^T \underline{p} = (1/6)n(n+1)(2n+1)$. Hence the lower bound in the theorem above is zero and the upper bound is $(1/6)\lambda_n(Q)n(n+1)(2n+1)$.

5.2. Projection bounds. Stronger bounds can be obtained by a projection technique described by Hadley, Rendl, and Wolkowicz [19]. The idea here is to satisfy both the stochasticity and orthonormality constraints and relax only the nonnegativity constraints. This technique involves projecting a permutation matrix X into a subspace orthogonal to the stochasticity constraints (5.1) by means of an eigenprojection.

Let the $n \times n-1$ matrix V be an orthonormal basis for the orthogonal complement of $\underline{u}$. By the choice of V, it satisfies two properties: $V^T\underline{u} = \underline{0}$ and $P = \begin{pmatrix} \underline{u} & V \end{pmatrix}$ is an orthonormal matrix of order n.

Observe that

$$P^T X P = \begin{pmatrix} \underline{u}^T \\ V^T \end{pmatrix} X \begin{pmatrix} \underline{u} & V \end{pmatrix} = \begin{pmatrix} \underline{u}^T X \underline{u} & \underline{u}^T XV \\ V^T X \underline{u} & V^T XV \end{pmatrix} = \begin{pmatrix} 1 & \underline{0}^T \\ \underline{0} & Y \end{pmatrix},$$

where $Y \equiv V^T XV$.

This suggests that we take

$$\begin{aligned} X &= P \begin{pmatrix} 1 & \underline{0}^T \\ \underline{0} & Y \end{pmatrix} P^T \\ &= \underline{u}\,\underline{u}^T + VYV^T. \end{aligned} \tag{5.4}$$

Note that with this choice the stochasticity constraints $X\underline{u} = \underline{u}$ and $X^T\underline{u} = \underline{u}$ are satisfied. Furthermore, if X is an orthonormal matrix of order n satisfying $X\underline{u} = \underline{u}$, then

$$P^T X P = \begin{pmatrix} 1 & \underline{0}^T \\ \underline{0} & Y \end{pmatrix}$$

is orthonormal, and this implies that Y is an orthonormal matrix of order $n-1$. Conversely, if Y is orthonormal of order $n-1$, then the matrix X obtained by the construction above is orthonormal of order n. The nonnegativity constraint $X \ge 0$ becomes, from (5.4), $VYV^T \ge -\underline{u}\,\underline{u}^T$. These facts will enable us to express the original QAP in terms of a projected QAP in the matrix of variables Y.

To obtain the projected QAP, we substitute the representation of X from (5.4) into the objective function $\operatorname{tr} QXBX^T$. Since $Q\underline{u} = \underline{0}$ by the construction of the Laplacian, terms of the form $Q\underline{u}\,\underline{u}^T\cdots$ vanish. Further,

$$\operatorname{tr} QVYV^TB\underline{u}\,\underline{u}^T = \operatorname{tr} \underline{u}^TQVYV^TB\underline{u},$$

where we use the identity $\operatorname{tr} MN = \operatorname{tr} NM$ for an $n\times k$ matrix M and a $k\times n$ matrix N. Again this term is zero since $\underline{u}^TQ = \underline{0}^T$. Hence the only nonzero term in the objective function is

$$\begin{aligned}&\operatorname{tr} Q\,VYV^T\,B\,VY^TV^T\\ &= \operatorname{tr}\,(V^TQV)\,Y\,(V^TBV)\,Y^T\\ &= \operatorname{tr}\,\widehat{Q}Y\widehat{B}Y^T,\end{aligned}$$

where $\widehat{M} = V^TMV$ is a projection of a matrix M.

We have obtained the projected QAP in terms of the matrix Y of order $n-1$, where the constraint that X be a permutation matrix now imposes the constraints that Y be orthonormal and that $VYV^T \geq -\underline{u}\underline{u}^T$. We obtain lower and upper bounds in terms of the eigenvalues of the matrices $\widehat{Q}$ and $\widehat{B}$ by relaxing the nonnegativity constraint again.

THEOREM 5.2. *The following upper and lower bounds hold for the* 2*-sum problem:*

$$(1/12)\lambda_2(Q)(n-1)n(n+1) \leq \sigma_2^2(A) \leq (1/12)\lambda_n(Q)(n-1)n(n+1).$$

Proof. If we apply the orthogonal bounds to the projected QAP, we get

$$\sum_{i=1}^{n-1}\lambda_i(\widehat{Q})\lambda_{n-i}(\widehat{B}) \leq \sigma_2^2(A) \leq \sum_{i=1}^{n-1}\lambda_i(\widehat{Q})\lambda_i(\widehat{B}).$$

The vector $\underline{u}$ is the eigenvector of Q corresponding to the zero eigenvalue, and hence eigenvectors corresponding to higher Laplacian eigenvalues are orthogonal to it. Thus any such eigenvector $\underline{x}_j$ can be expressed as $\underline{x}_j = V\underline{r}_j$. Substituting this last equation into the eigenvalue equation $Q\underline{x}_j = \lambda_j(Q)\underline{x}_j$ and premultiplying by V^T, we obtain $\widehat{Q}\underline{r}_j = \lambda_j(Q)\underline{r}_j$. Hence for $i = 2, \ldots, n$, we have $\lambda_i(Q) = \lambda_{i-1}(\widehat{Q})$. Also, $\lambda_{n-1}(\widehat{B}) = \underline{p}^TVV^T\underline{p}$, and all other eigenvalues are zero. Hence it remains to compute the largest eigenvalue of $\widehat{B}$.

From the representation $I_n = PP^T = \underline{u}\,\underline{u}^T + VV^T$, we compute

$$\begin{aligned}&\underline{p}^TVV^T\underline{p}\\ &= \underline{p}^T\,\underline{p} - (\underline{p}^T\,\underline{u})\,(\underline{u}^T\,\underline{p})\\ &= (1/6)n(n+1)(2n+1) - (1/4)n(n+1)^2 \;=\; (1/12)(n-1)n(n+1).\end{aligned}$$

We get the result by substituting these eigenvalues into the bounds for the 2-sum. □

For later use in justifying the spectral algorithm for minimizing the 2-sum, we observe that the lower bound is attained by the matrix

$$X_0 = \underline{u}\,\underline{u}^T + VRS^TV^T, \tag{5.5}$$

where R (S) is a matrix of eigenvectors of $\widehat{Q}$ ($\widehat{B}$) and the eigenvectors correspond to the eigenvalues of $\widehat{Q}$ ($\widehat{B}$) in nondecreasing (nonincreasing) order.

The result given above has been obtained by Juvan and Mohar [24] without using a QAP formulation of the 2-sum. We have included this proof for two reasons. First, in the next subsection, we show how the lower bound may be strengthened by diagonal perturbations of the Laplacian. Second, in the following section, we consider the problem of finding a permutation matrix closest to the orthogonal matrix attaining the lower bound.

5.3. Diagonal perturbations. The lower bound for the 2-sum can be further improved by perturbing the Laplacian matrix Q by a diagonal matrix $\mathrm{Diag}(\underline{d})$, where $\underline{d}$ is an n-vector, and then using an optimization routine to maximize the smallest eigenvalue of the perturbed matrix.

Choosing the elements of $\underline{d}$ such that its elements sum to zero, i.e., $\underline{u}^T\underline{d} = 0$, simplifies the bounds we obtain, and hence we make this assumption in this subsection. We begin by denoting $Q(d) = Q + \mathrm{Diag}(\underline{d})$ and expressing

$$f(X) \equiv \mathrm{tr}\, QXBX^T = \mathrm{tr}\, Q(d)XBX^T - \mathrm{tr}\, \mathrm{Diag}(\underline{d})XBX^T.$$

The second term can be written as a linear assignment problem (LAP) since one of the matrices involved is diagonal. Let the permutation vector $\underline{\alpha} = X\underline{p}$, and let $\underline{d}_B$ denote the n-vector formed from the diagonal elements of B:

$$\mathrm{tr}\, \mathrm{Diag}(\underline{d})XBX^T = \sum_{i=1}^{n} d_i b_{\alpha(i),\alpha(i)} = \mathrm{tr}\, \underline{d}\, \underline{d}_B{}^T X^T.$$

We now proceed, as in the previous subsection, to obtain projected bounds for the quadratic term, and thus for $f(X)$. Note that $Q(d)\underline{u} = (1/\sqrt{n}\,)\underline{d}$ since $Q\underline{u} = \underline{0}$ and $\underline{u}^TQ(d)\underline{u} = 0$ since the elements of $\underline{d}$ sum to zero. We let $B\underline{u} = (1/\sqrt{n}\,)\,\underline{r}(B)$ denote the row sum of the elements of B.

With notation as in the previous subsection, we substitute $X = \underline{u}\,\underline{u}^T + VYV^T$ in the quadratic term in $f(X)$. The first term $\mathrm{tr}Q(d)\underline{u}\underline{u}^TB\underline{u}\underline{u}^T = \mathrm{tr}\underline{u}^TQ(d)\underline{u}\underline{u}^TB\underline{u} = 0$. The second and third terms are equal, and their sum can be transformed as follows:

$$\begin{aligned}
&2\,\mathrm{tr}\, Q(d)VYV^TB\underline{u}\,\underline{u}^T = 2\,\mathrm{tr}\, \underline{u}^TQ(d)VYV^TB\underline{u} \\
&= (2/n)\,\mathrm{tr}\, \underline{d}^TVYV^T\underline{r}(B) = (2/n)\,\mathrm{tr}\, V^T\underline{r}(B)\,\underline{d}^TVY \\
&= (2/n)\,\mathrm{tr}\, Y^TV^T\underline{d}\,\underline{r}(B)^TV = (2/n)\,\mathrm{tr}\, \underline{d}\,\underline{r}(B)^TVY^TV^T.
\end{aligned}$$

Note that this term is linear in the projected variables Y, and we shall find it convenient to express it in terms of X by the substitution $X^T - \underline{u}\,\underline{u}^T = VY^TV^T$. Thus

$$(2/n)\,\mathrm{tr}\, \underline{d}\,\underline{r}(B)^TVY^TV^T = (2/n)\,\mathrm{tr}\, \underline{d}\,\underline{r}(B)^T(X^T - \underline{u}\,\underline{u}^T) = (2/n)\,\mathrm{tr}\, \underline{d}\,\underline{r}(B)^TX^T$$

since the second term is equal to $\mathrm{tr}\, \underline{u}^T\,\underline{d}\,\underline{r}(B)^T\underline{u}$, which is zero by the choice of $\underline{d}$.

Finally, the fourth term becomes $\mathrm{tr}\,\widehat{Q}(d)Y\widehat{B}Y^T$, where $\widehat{Q}(d) = V^TQ(d)V$, and as before $\widehat{B} = V^TBV$.

Putting it all together, we obtain

$$f(X) = \mathrm{tr}\,\widehat{Q}(d)Y\widehat{B}Y^T + \mathrm{tr}\,\left((2/n)\underline{d}\,\underline{r}(B)^TX^T - \underline{d}\,\underline{d}_B{}^TX^T\right).$$

Observe that the first term is quadratic in the projected variables Y and the remaining terms are linear in the original variables X. Our lower bound for the 2-sum shall be obtained by minimizing the quadratic and linear terms separately.

We can simplify the LAP by noting that $B = \underline{p}\,\underline{p}^T$. Thus $r_{B,i} = i\sum_{j=1}^{n} j = (1/2)n(n+1)i$, and hence $(2/n)\underline{r}(B) = (n+1)\underline{p}$. Further, $\underline{d}_B = sq(\underline{p})$, the vector with ith component equal to i^2. Hence the final expression for the LAP is

$$\operatorname{tr} \underline{d}\left((n+1)\underline{p}^T - sq(\underline{p})^T\right) X^T.$$

The minimum value of this problem, denoted by $L(\underline{d})$ (the minimum over the permutation matrices X for a given $\underline{d}$), can be computed by sorting the components of $\underline{d}$ and $\big((n+1)\underline{p} - sq(\underline{p})\big)$.

The eigenvalues of $\widehat{B}$ can be computed as in the previous subsection. We may choose $\underline{d}$ to maximize the lower bound. Thus this discussion leads to the following result.

THEOREM 5.3. *The minimum* 2*-sum of a symmetric matrix A can be bounded as*

$$\sigma^2_{2,\,\min}(A) \geq \max_{\underline{d}} \left\{(1/12)\lambda_1(\widehat{Q}(d))(n-1)n(n+1) + L(\underline{d})\right\},$$

where the components of the vector $\underline{d}$ *sum to zero.* □

6. Computing an approximate solution from the lower bound. Consider the problem of finding a permutation matrix Z "closest" to an orthogonal matrix X_0 that attains the lower bound in Theorem 5.2. We show in this section that sorting the second Laplacian eigenvector components in nonincreasing (also nondecreasing) order yields a permutation matrix that solves a linear approximation to the problem. This justifies the spectral approach for minimizing the 2-sum.

From (5.5), the orthogonal matrix $X_0 = \underline{u}\,\underline{u}^T + VRS^TV^T$, where R (S) is a matrix of eigenvectors of $\widehat{Q}$ ($\widehat{B}$) corresponding to the eigenvalues of $\widehat{Q}$ ($\widehat{B}$) in increasing (decreasing) order. We begin with a preliminary discussion of some properties of the matrix X_0 and the eigenvectors of Q. For $j = 1, \ldots, n-1$, let the jth column of R be denoted by $\underline{r}_j$, and similarly let $\underline{s}_j$ denote the jth column of S. Then $\underline{s}_1 = cV^T\underline{p}$, where c is a normalization constant; for $j = 2, \ldots, n-1$, the vector $\underline{s}_j$ is orthogonal to $V^T\underline{p}$, i.e.,

$$\underline{s}_j{}^T V^T \underline{p} = 0. \tag{6.1}$$

Recall from the previous section that a second Laplacian eigenvector $\underline{x}_2 = V\underline{r}_1$.

Now we can formulate the "closest" permutation matrix problem more precisely. The minimum 2-sum problem may be written as

$$\min_{Z} \|(Q+\alpha I)^{1/2} Z\underline{p}\|_2{}^2.$$

We have chosen a positive shift α to make the shifted matrix positive definite and hence to obtain a weighted norm by making the square root nonsingular. It can be verified that the shift has no effect on the minimizer since it adds only a constant term to the objective function.

We substitute $Z = X_0 + (Z - X_0)$ and expand the 2-sum about X_0 to obtain

(6.2)

$$\begin{aligned}&\|(Q+\alpha I)^{1/2} Z\underline{p}\|_2{}^2 \\ &= \|(Q+\alpha I)^{1/2} X_0\,\underline{p}\|_2{}^2 + 2\operatorname{tr}\underline{p}^T(Z-X_0)^T(Q+\alpha I)X_0\,\underline{p} + \|(Q+\alpha I)^{1/2}(Z-X_0)\,\underline{p}\|_2{}^2.\end{aligned}$$

The first term on the right-hand side is a constant since X_0 is a given orthogonal matrix; the third term is a quadratic in the difference $(Z - X_0)$ and hence we neglect it to obtain a linear approximation. It follows that we can choose a permutation matrix Z close to X_0 to approximately minimize the 2-sum by solving

$$\min_Z \operatorname{tr} \underline{p}^T Z^T (Q + \alpha I) X_0 \underline{p} = \min_Z \operatorname{tr} (Q + \alpha I) X_0 B Z^T. \tag{6.3}$$

Substituting for X_0 from (5.5) in this LAP and noting that $Q\underline{u} = \underline{0}$, we find

$$\begin{aligned} \min_Z \operatorname{tr} (Q + \alpha I) X_0 B Z^T &= \min_Z \operatorname{tr} (Q + \alpha I) (\underline{u}\, \underline{u}^T + V R S^T V^T) B Z^T \\ &= \min_Z \left(\operatorname{tr} Q V R S^T V^T B Z^T + \alpha \operatorname{tr} \underline{u}\, \underline{u}^T B Z^T + \alpha \operatorname{tr} V R S^T V^T B Z^T \right). \end{aligned} \tag{6.4}$$

The second term on the right-hand side is a constant since

$$\operatorname{tr} \underline{u}\, \underline{u}^T B Z^T = \operatorname{tr} \underline{u}^T B Z^T \underline{u} = \operatorname{tr} \underline{u}^T B \underline{u} = (\underline{u}^T\, \underline{p})^2.$$

Here we have substituted $Z^T \underline{u} = \underline{u}$ from (5.1). We proceed to simplify the first term in (6.4), which is

$$\operatorname{tr} Q V R S^T V^T B Z^T = \operatorname{tr} Q V \left(\sum_{j=1}^{n-1} \underline{r}_j\, \underline{s}_j{}^T \right) V^T \underline{p}\, \underline{p}^T Z^T.$$

From (6.1) we find that $\underline{s}_j{}^T V^T \underline{p} = 0$, for $j = 2, \ldots, n-1$, and hence *only the first term in the sum survives*. Noting that $\underline{s}_1 = c V^T \underline{p}$ and $V \underline{r}_1 = \underline{x}_2$, this term becomes

$$\operatorname{tr} Q \underline{x}_2\, (c \underline{p}^T V)\, V^T \underline{p}\, \underline{p}^T Z^T = c \lambda_2(Q) (\underline{p}^T V V^T \underline{p}) \operatorname{tr} \underline{x}_2\, \underline{p}^T Z^T.$$

The third term in (6.4) can be simplified in a similar manner, and hence ignoring the constant second term, this equation becomes

$$c(\lambda_2(Q) + \alpha)\, (\underline{p}^T V V^T \underline{p}) \min_Z \operatorname{tr} \underline{x}_2\, \underline{p}^T Z^T.$$

Hence we are required to choose a permutation matrix Z to minimize $\operatorname{tr} \underline{x}_2 \underline{p}^T Z^T = \operatorname{tr} Z^T \underline{x}_2 \underline{p}^T$. The solution to this problem is to choose Z to correspond to a permutation of the components of $\underline{x}_2$ in nonincreasing order, since the components of the vector $\underline{p}$ are in increasing order. Note that $-\underline{x}_2$ is also an eigenvector of the Laplacian matrix, and since the positive or negative signs of the components are chosen arbitrarily, sorting the eigenvector components into nondecreasing order also gives a permutation matrix Z closest, within a linear approximation, to a different choice for the orthogonal matrix X_0 (see (5.5)).

Similar techniques can be used to show that if one is interested in *maximizing* the 2-sum, then a closest permutation matrix to the orthogonal matrix that attains the upper bound in Theorem 5.2 is approximated by sorting the components of the Laplacian eigenvector $\underline{x}_n$ (corresponding to the largest eigenvalue $\lambda_n(Q)$) in nondecreasing (nonincreasing) order.

7. Asymptotic behavior of envelope parameters. In this section, we first prove that graphs with good separators have asymptotically small envelope parameters and next study the asymptotic behavior of the lower bounds on the envelope parameters as a function of the problem size.

7.1. Upper bounds on envelope parameters. Let α, β, and γ be constants such that $(1/2) \leq \alpha, \gamma < 1$, and define $n_0 \equiv (\beta/(1-\alpha))^{1/(1-\gamma)}$. A class of graphs $\mathcal{G}$ has n^γ-separators if every graph G on $n > n_0$ vertices in $\mathcal{G}$ can be partitioned into three sets A, B, S such that no vertex in A is adjacent to any vertex in B, and the number of vertices in the sets are bounded by the relations $|A|, |B| \leq \alpha n$ and $|S| \leq \beta n^\gamma$. If $n \leq n_0$, then we choose the separator S to consist of the entire graph. If $n > n_0$, then by the choice of n_0,

$$\alpha n + \beta n^\gamma = n\left(\alpha + \beta n^{\gamma-1}\right) < n\left(\alpha + \beta n_0^{\gamma-1}\right) = n,$$

and we separate the graph into two parts A and B by means of a separator S. The assumption that γ is at least a half is not a restriction for the classes of graphs that we are interested in here: planar graphs have $n^{1/2}$-separators and overlap graphs [30] embedded in $d \geq 2$ dimensions have $n^{(d-1)/d}$-separators. The latter class includes "well-shaped" finite element graphs in d dimensions, i.e., finite element graphs with elements of bounded aspect ratio.

THEOREM 7.1. *Let $\mathcal{G}$ be a class of graphs that has n^γ-separators and maximum vertex degree bounded by Δ. The minimum envelope size* $\mathrm{Esize}_{\min}(G)$ *of any graph $G \in \mathcal{G}$ on n vertices is $\mathcal{O}(n^{1+\gamma})$.*

Proof. If $n \leq n_0$, then we order the vertices of G arbitrarily. Otherwise, let a separator S separate G into the two sets A and B, where we choose the subset B to have no more vertices than A. We consider a "modified nested dissection" ordering of G that orders the vertices in A first, the vertices in S next, and the vertices in B last. (See the ordering in Figure 2.1, where S corresponds to the set of vertices in the middle column.)

The contribution to the envelope E_S made by the vertices in S is bounded by the product of the maximum row width of a vertex in S and the number of vertices in S. Thus

$$E_S \leq |S| \cdot |A \cup S| \leq \beta n^\gamma(\alpha n + \beta n^\gamma) = \alpha\beta n^{1+\gamma} + \beta^2 n^{2\gamma}.$$

We also consider the contribution made by vertices in B that are adjacent to nodes in S as a consequence of numbering the nodes in S. There are at most $\Delta|S|$ such vertices in B. Since these vertices are not adjacent to any vertex in A, the contribution E_B made by them is

$$E_B \leq \Delta|S| \cdot |B \cup S| \leq \Delta\beta n^\gamma(\alpha n + \beta n^\gamma) = \Delta\alpha\beta n^{1+\gamma} + \Delta\beta^2 n^{2\gamma}.$$

Let n_1 (n_2) denote the number of vertices in the subset A (B). Adding the contributions from the two sets of nodes in the previous paragraph, we obtain the recurrence relation

$$E(n) \leq \alpha\beta(1+\Delta)n^{1+\gamma} + \beta^2(1+\Delta)n^{2\gamma} + \max_{n_1,n_2}\left(E(n_1) + E(n_2)\right), \tag{7.1}$$

where $n_1, n_2 \leq \alpha n$ and $n_1 + n_2 \leq n$.

We claim that

$$E(n) \leq C_1 n^{1+\gamma} + C_2 n^{2\gamma}\log n \tag{7.2}$$

for suitable constants C_1 and C_2 to be chosen later. We prove the claim by induction on n.

For $n \leq n_0$, the claim may be satisfied by choosing C_1 to be greater than or equal to $(n_0+1)/2$, since

$$E(n) \leq n(n+1)/2 \leq n(n_0+1)/2 \leq C_1 n^{1+\gamma}.$$

Now consider the case when $n > n_0$. Let the maximum in the recurrence relation (7.1) be attained for $n_1 = an$ and $n_2 = bn \leq (1-a)n$, where $1/2 \leq a \leq \alpha < 1$. Since $n > n_0$, we have $n_1, n_2 < n$; thus the inductive hypothesis can be applied to the subgraphs induced by A and B. Hence we substitute the bound (7.2) into recurrence relation (7.1) to obtain

$$\begin{aligned} E(n) \leq & \left(\alpha\beta(1+\Delta) + C_1(a^{1+\gamma} + (1-a)^{1+\gamma})\right) n^{1+\gamma} \\ & + \left(\beta^2(1+\Delta) + C_2(a^{2\gamma}\log an + (1-a)^{2\gamma}\log(1-a)n)\right) n^{2\gamma}. \end{aligned}$$

For the claim to be satisfied, this bound must be less than the right-hand side of inequality (7.2). We prove this by considering the coefficients of the terms $n^{1+\gamma}$ and $n^{2\gamma}$.

Consider the $n^{1+\gamma}$ term first. It is easy to see that $a^{1+\gamma} + (1-a)^{1+\gamma} < 1$ because $1/2 \leq a \leq \alpha < 1$ and γ is positive. Furthermore, this expression attains its maximum when a is equal to α. Denote this maximum value by $\epsilon \equiv \alpha^{1+\gamma} + (1-\alpha)^{1+\gamma} < 1$. Equating the coefficients of $n^{1+\gamma}$ in the recurrence relation, if

$$C_1\epsilon + \alpha\beta(1+\Delta) \leq C_1,$$

then the first term in the claimed asymptotic bound on $E(n)$ would be true. Both this inequality and the condition on C_1 imposed by n_0 are satisfied if we choose

$$C_1 \geq \max\left\{\frac{\alpha\beta(1+\Delta)}{1-\epsilon}, (n_0+1)/2\right\}.$$

We simplify the coefficient of the $n^{2\gamma}$ term a bit before proceeding to analyze it. We have

$$\begin{aligned} & a^{2\gamma}\log an + (1-a)^{2\gamma}\log(1-a)n \\ & \leq a^{2\gamma}\log an + (1-a)^{2\gamma}\log an \;\leq\; \left(\alpha^{2\gamma} + (1-\alpha)^{2\gamma}\right)\log_\alpha n \equiv \theta\log_\alpha n \\ & \leq \log_\alpha n. \end{aligned}$$

In the transformations we have used the following facts: $1-a \leq a$ since $a \geq 1/2$; the maximum of $a^{2\gamma} + (1-a)^{2\gamma}$, when $1/2 \leq a \leq \alpha$ and 2γ is greater than or equal to one, is attained for $a = \alpha$; this maximum value θ is less than one. Hence for the claim to hold, we require

$$C_2\log_\alpha n + \beta^2(1+\Delta) \leq C_2\log n.$$

This last inequality is satisfied if we choose

$$C_2 \geq \frac{\beta^2(1+\Delta)}{\log\alpha^{-1}}. \qquad \square$$

A similar proof yields $\text{Wbound}_{\min}(G) = \mathcal{O}(n^{2+\gamma})$, which is an upper bound on the work in an envelope Cholesky factorization. Hence good separators imply small envelope size and work. Although we have used a "modified nested dissection" ordering to prove asymptotic upper bounds, we do not advocate the use of this ordering for envelope reduction. Other envelope reducing algorithms considered in this paper are preferable because they are faster and yield smaller envelope parameters.

TABLE 7.1
Asymptotic upper and lower bounds on envelope size and work for an overlap graph in d dimensions.

problem	separator size	λ_2	Esize(A)		Ework(A)	
			LB	UB	LB	UB
d-dim.	$\mathcal{O}(n^{1-1/d})$	$\Theta(n^{-2/d})$	$\Omega(n^{2-2/d})$	$\mathcal{O}(n^{2-1/d})$	$\Omega(n^{3-4/d})$	$\mathcal{O}(n^{3-1/d})$

7.2. Asymptotic behavior of lower bounds. In this subsection we consider the implications of the spectral lower bounds that we have obtained. We denote the eigenvalue $\lambda_2(Q)$ by λ_2 for the sake of brevity in this subsection. We use the asymptotic behavior of the second eigenvalues together with the lower bounds we have obtained to predict the behavior of envelope parameters. For the envelope size, we make use of Theorem 3.2; for the envelope work, we employ Theorem 3.3.

The bounds on envelope parameters are tight for dense and random graphs (matrices). For instance, the full matrix (the complete graph) has $\lambda_2 = \Delta + 1 = n$, and hence $\text{Esize}_{\min}(A) = \Theta(n^2)$. Similarly, the bound on the envelope work $\text{Ework}_{\min}(A) = \Theta(n^3)$. The predicted lower bound is within a factor of three of the envelope size. These bounds are also asymptotically tight for random graphs where each possible edge is present in the graph with a given constant probability p, since the second Laplacian eigenvalue satisfies [23]

$$\lambda_2 = pn - \Theta([p(1-p)n \log n]^{1/2}).$$

More interesting are the implications of these bounds for degree-bounded finite element meshes in two and three dimensions. We will employ the following result proved recently by Spielman and Teng [38].

THEOREM 7.2. *The second Laplacian eigenvalue of an overlap graph embedded in d dimensions is bounded by* $\mathcal{O}(n^{-2/d})$. □

Planar graphs are overlap graphs in two dimensions and well-shaped meshes in three dimensions are also overlap graphs with $d = 3$.

Table 7.1 summarizes the asymptotic lower and upper bounds on the envelope parameters for a well-shaped mesh embedded in d dimensions. The most useful values are $d = 2$ and $d = 3$. As before, the lower bound on the envelope size is from Theorem 3.2, while the lower bound on the envelope work is from Theorem 3.3. The upper bound on the envelope size follows from Theorem 7.1, and the upper bound on the envelope work follows from the upper bound on Wbound(A), discussed at the end of the proof of that theorem.

The lower bounds are obtained *for problems where the upper bounds on the second eigenvalue are asymptotically tight.* This is reasonable for many problems, for instance, model problems in partial differential equations. Note that the regular finite element mesh in a discretization of Laplace's equation in two dimensions (Neumann boundary conditions) has $\lambda_2 = \Theta(h^2) = \Theta(n^{-1})$, where h is the smallest diameter of an element (smallest mesh spacing for a finite difference mesh). The regular three-dimensional mesh in the discretized Laplace's equation with Neumann boundary conditions satisfies $\lambda_2 = \Theta(h^2) = \Theta(n^{-2/3})$.

For planar problems, the lower bound on the envelope size is $\Omega(n)$, while the upper bound is $\mathcal{O}(n^{1.5})$. For well-shaped three-dimensional meshes, these bounds are $\Omega(n^{4/3})$ and $\mathcal{O}(n^{5/3})$. The lower bounds on the envelope work are weaker since they

TABLE 8.1
2-sums from the spectral reordering algorithm and lower bounds for triangulations of the sphere.

$\|V\|$	$\|E\|$	λ_2	Spectral LB	Spectral 2-sum	Gap(%)
18	48	2.00	969	978	0.9
66	192	6.25e-1	1.50e+4	1.54e+4	2.6
258	768	1.65e-1	2.36e+5	2.53e+5	6.9
1,026	3,072	4.17e-2	3.75e+6	4.05e+6	7.4
4,098	12,270	1.05e-2	6.00e+7	6.44e+7	7.3
16,386	49,152	2.60e-3	0.953e+9	1.03e+9	9.1

are obtained from the corresponding bounds on the envelope size. Direct methods for solving sparse systems have storage requirements bounded by $\mathcal{O}(n \log n)$ and work bounded by $\mathcal{O}(n^{1.5})$ for a two-dimensional mesh; in well-shaped three-dimensional meshes these are $\mathcal{O}(n^{4/3})$ and $\mathcal{O}(n^2)$.

These results suggest that when a two-dimensional mesh possesses a small second Laplacian eigenvalue, envelope methods may be expected to work well. Similar conclusions should hold for three-dimensional problems when the number of mesh points along the third dimension is small relative to the number in the other two dimensions and for two-dimensional surfaces embedded in three-dimensional space.

8. Computational results. We now present computational results to verify how well the spectral ordering reduces the 2-sum. We report results on two sets of problems.

The first set of problems, shown in Table 8.1, is obtained from John Richardson's (Thinking Machines Corporation) program for triangulating the sphere. The spectral lower bounds reported are from Theorem 5.2. Gap is the ratio with numerator equal to the difference between the 2-sum and the lower bound and the denominator equal to the 2-sum. The results show that the spectral reordering algorithm computes values within a few percent of the optimal 2-sum, since the gap between the spectral 2-sum and the lower bound is within that range.

Table 8.2 contains the second set of problems, taken from the Boeing–Harwell and NASA collections. Here the bounds are weaker than those in Table 8.1. These problems have two features that distinguish them from the sphere problems. Many of them have less regular degree distributions; e.g., NASA1824 has maximum degree 41 and minimum degree 5. They also represent more complex geometries. Nevertheless, these results imply that the spectral 2-sum is within a factor of two of the optimal value for these problems. These results are somewhat surprising since we have shown that minimizing the 2-sum is NP-complete.

The gap between the computed 2-sums and the lower bounds could be further reduced in two ways. First, a local reordering algorithm applied to the ordering computed by the spectral algorithm might potentially decrease the 2-sum. Second, the lower bounds could be improved by incorporating diagonal perturbations to the Laplacian.

9. Conclusions. The lower bounds on the 2-sums show that the spectral reordering algorithm can yield nearly optimal values in spite of the fact that minimizing the 2-sum is an NP-complete problem. To the best of our knowledge, these are the first results providing reasonable bounds on the quality of the orderings generated by a reordering algorithm for minimizing envelope-related parameters. Earlier work had not addressed the issue of the quality of the orderings generated by the algorithms.

TABLE 8.2
2-sums from the spectral reordering algorithm and lower bounds for some problems from the Boeing–Harwell and NASA collections.

Problem	$\|V\|$	$\|E\|$	λ_2	Spectral LB	Spectral 2-sum	Gap(%)
CAN1072	1,072	5,686	7.96e–2	8.17e+6	9.02e+6	9.4
NASA1824	1,824	18,692	2.71e–1	1.37e+8	1.74e+8	21
NASA2146	2,146	35,052	1.35e–1	1.11e+8	1.32e+8	16
NACA	4,224	12,416	3.57e–3	2.24e+7	2.70e+7	17
BARTH4	6,019	17,473	1.76–3	3.19e+7	5.41e+7	41
BARTH	6,691	19,748	2.62e–3	6.54e+7	6.69e+7	2.2
BARTH5	15,606	45,878	7.41e–4	2.35e+8	3.06e+8	23
BCSSTK30	28,924	1,007,284	1.96e–2	3.00e+10	5.73e+10	48
COPTER2	55,476	352,238	6.77e–3	9.63e+10	1.17e+11	18

Unfortunately, the tight bounds on the 2-sum do not lead to tight bounds on the envelope parameters. However, we have shown that problems with bounded separator sizes have bounded envelope parameters and have obtained asymptotic lower and upper bounds on these parameters for finite element meshes.

Our analysis further shows that the spectral orderings attempt to minimize the 2-sum rather than the envelope parameters. Hence a reordering algorithm could be used in a postprocessing step to improve the envelope and wavefront parameters from a spectral ordering. A combinatorial reordering algorithm called the Sloan algorithm has been recently used by Kumfert and Pothen [25] to reduce envelope size and front widths. Currently this algorithm computes the lowest values of the envelope parameters on a collection of finite element meshes.

Acknowledgments. Professor Stan Eisenstat (Yale University) carefully read two drafts of this paper and pointed out several errors. Every author should be so blessed! Thanks, Stan.

Appendix. Lower bounds on the minimum *p*-sum. We prove two lower bounds on the minimum p-sum. We make use of Lemma 3.1 in proving the first result. In the following $B_m(x)$ is the mth Bernoulli polynomial and B_m is the mth Bernoulli number.

THEOREM A.1. *For $1 \le p < \infty$, the minimum p-sum of a graph G on n vertices satisfies*

$$\sigma^p_{p,\,\min}(G) \ge \frac{1}{p+1}\left(B_{p+1}(s+1) - B_{p+1}\right),$$

where $s = (\lambda_2/4\Delta)n$.

Proof. Consider any ordering α of the vertices of G. Partition the vertices into two sets: A consisting of the lowest numbered $n/2$ vertices and B consisting of the highest numbered $n/2$ vertices. By Lemma 3.1 the number of edges joining A and B, $|\delta(A,B)|$, is

$$|\delta(A,B)| \ge \frac{\lambda_2}{n}(n/2)^2.$$

Hence at least $s = |\delta(A,B)|/\Delta$ vertices in B are adjacent to vertices in A. Each vertex in this subset of B has the smallest row width when it is adjacent to the highest numbered vertex in A and to no other vertices in A. Hence these s vertices

make a contribution of at least $1^p + \cdots + s^p$ to the p-sum, and this sum can be expressed in terms of the Bernoulli polynomials, as stated. □

From an expansion of the Bernoulli polynomial, we find that asymptotically

$$\sigma^p_{p,\,\min}(G) \geq \frac{1}{(p+1)(4\Delta)^{p+1}} {\lambda_2}^{p+1} n^{p+1} + \mathcal{O}(({\lambda_2}^p/\Delta^p) n^p).$$

We proceed to obtain another lower bound on the minimum p-sum.

The next result makes use of the following lemma recently proved by Helmberg et al. [20]. Define the following symmetric function of the two positive integers m_1, m_2 (with $m_1 + m_2 < n$) and parameters λ_2, λ_n:

$$f(m_1, m_2) = \frac{\sqrt{m_1 m_2}}{2n} \left[\left(\sqrt{m_1 m_2} + \sqrt{(n-m_1)(n-m_2)} \right) \lambda_2 + \left(\sqrt{m_1 m_2} - \sqrt{(n-m_1)(n-m_2)} \right) \lambda_n \right]. \tag{A.1}$$

LEMMA A.2. *Let S_1, S_2 be two disjoint subsets of the vertices of a graph G on n vertices, with $|S_i| = s_i$, for $i = 1, 2$. Then the number of edges joining S_1 and S_2, $|\delta(S_1, S_2)|$, satisfies*

$$|\delta(S_1, S_2)| \geq f(s_1, s_2). \qquad \square$$

THEOREM A.3. *For $1 \leq p < \infty$, the minimum p-sum of a graph G satisfies*

$$\sigma^p_{p,\,\min}(G) \geq \frac{1}{2^{p+1}\Delta} \frac{{\lambda_2}^{p+1}}{(\lambda_n + \lambda_2)^{p+2}} (2\lambda_n + \lambda_2)(\lambda_n + 2\lambda_2) n^{p+1}.$$

Proof. Consider any ordering α of the vertices of G and a tripartition A, B, C: we choose A to consist of the lowest numbered $a \equiv (n-b)/2$ vertices, C to consist of the highest numbered $(n-b)/2$ vertices, and B to contain the remaining b vertices in the "middle." Here b, the size of B, is a parameter that will be determined later to obtain a large lower bound.

From Lemma A.2, $|\delta(A, C)|$, the number of edges joining A and C, is at least $f(a, a)$, where the symmetric function $f(.,.)$ is defined in (A.1). Hence there are at least $s_C = f(a, a)/\Delta$ vertices in C adjacent to vertices in A. Each of these vertices has row width at least b.

Initially, consider the contribution to the envelope size $\mathrm{Esize}(G)$ made by these vertices to obtain a suitable value for b:

$$\begin{aligned} \mathrm{Esize}(G) &\geq \frac{f(a,a)}{\Delta} b \\ &= \frac{(n-b)}{4n} \left[\left(\frac{n-b}{2} + \frac{n+b}{2} \right) \lambda_2 + \left(\frac{n-b}{2} - \frac{n+b}{2} \right) \lambda_n \right] \frac{b}{\Delta} \\ &= \frac{1}{4\Delta} b(n-b) \left(\lambda_2 - (b/n)\lambda_n \right). \end{aligned} \tag{A.2}$$

We choose b to maximize the lower bound on the envelope size. Differentiating the cubic polynomial in (A.2) with respect to b and simplifying, we obtain the quadratic equation

$$b^2 - \frac{2}{3} \frac{\lambda_2 + \lambda_n}{\lambda_n} nb + \frac{1}{3} \frac{\lambda_2}{\lambda_n} n^2 = 0.$$

From the quadratic we find that the maximizer is, to first order, $b_m = (1/2)(\lambda_2/(\lambda_n + \lambda_2))n$.

Now we consider the contribution to the p-sum made by the s_C vertices in C adjacent to vertices in A. Each of these vertices contributes at least b^p to the p-sum, and thus a lower bound on the minimum p-sum is

$$\sigma^p_{p,\min}(G) \geq \frac{1}{4\Delta}(n-b)\,(\lambda_2 - (b/n)\lambda_n)\,b^p.$$

It is not easy to find a maximizer of the right-hand side in the bound above on the p-sum since the polynomial in b is of degree $p+2$. Hence we choose b equal to the maximizer of the envelope size. We obtain the bound stated in the theorem by substituting $b = b_m$ in the bound above. □

Juvan and Mohar [24] have proved upper bounds for the p-sums. The techniques in this Appendix can be used to compute bounds on Esize(A) and Wbound(A), but the results are weaker than those obtained in section 3.

REFERENCES

[1] N. Alon and V. Milman, λ_1, *isoperimetric inequalities for graphs, and superconcentrators*, J. Combin. Theory Ser. B, 38 (1985), pp. 73–88.

[2] S. T. Barnard, A. Pothen, and H. D. Simon, *A spectral algorithm for envelope reduction of sparse matrices*, Numer. Linear Algebra Appl., 2 (1995), pp. 317–334. (A shorter version appeared in Supercomputing '93, IEEE Computer Society Press, 1993, pp. 493–502.)

[3] E. H. Cuthill and J. McKee, *Reducing the bandwidth of sparse symmetric matrices*, in Proc. 24th Nat. Conf. Assoc. Comp. Mach., ACM Publications, 1969, pp. 157–172.

[4] E. F. D'Azevedo, P. A. Forsyth, and W. P. Tang, *Ordering methods for preconditioned conjugate gradient methods applied to unstructured grid problems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 944–961.

[5] I. S. Duff, A. M. Erisman, and J. K. Reid, *Direct Methods for Sparse Matrices*, Clarendon Press, Oxford, UK, 1986.

[6] I. S. Duff and G. A. Meurant, *The effect of ordering on preconditioned conjugate gradients*, BIT, 29 (1989), pp. 635–657.

[7] I. S. Duff, J. K. Reid, and J. A. Scott, *The use of profile reduction algorithms with a frontal code*, Internat. J. Numer. Methods Engrg., 28 (1989), pp. 2555–2568.

[8] S. Even, *Graph Algorithms*, Computer Science Press, Rockville, MD, 1979.

[9] M. Fiedler, *Algebraic connectivity of graphs*, Czech. Math. J., 23 (1973), pp. 298–305.

[10] M. Fiedler, *A property of eigenvectors of non-negative symmetric matrices and its application to graph theory*, Czech. Math. J., 25 (1975), pp. 619–633.

[11] G. Finke, R. F. Burkard, and F. Rendl, *Quadratic assignment problems*, in Surveys in Combinatorial Optimization, S. Martell et al., eds., Annals of Discrete Mathematics, Vol. 31, Elsevier Science Publishers, New York, 1987, pp. 61–82.

[12] N. Gaffke and O. Krafft, *Matrix inequalities in the Löwner orderings*, in Modern Applied Mathematics: Optimization and Operations Research, B. Korte, ed., North–Holland, Amsterdam, 1982, pp. 576–622.

[13] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, CA, 1979.

[14] A. George, *Computer implementation of the Finite Element Method*, Tech. report 208, Department of Computer Science, Stanford University, Stanford, CA, 1971.

[15] A. George and J. W. H. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1981.

[16] N. E. Gibbs, *Algorithm* 509 : *A hybrid profile reduction algorithm*, ACM Trans. Math. Software, 2 (1976), pp. 378–387.

[17] N. E. Gibbs, W. G. Poole, Jr., and P. K. Stockmeyer, *An algorithm for reducing the bandwidth and profile of a sparse matrix*, SIAM J. Numer. Anal., 13 (1976), pp. 236–249.

[18] R. G. Grimes, D. J. Pierce, and H. D. Simon, *A new algorithm for finding a pseudoperipheral node in a graph*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 323–334.

[19] S. Hadley, F. Rendl, and H. Wolkowicz, *A new lower bound via projection for the quadratic assignment problem*, Math. Oper. Res., 17 (1992), pp. 727–739.

[20] C. Helmberg, B. Mohar, S. Poljak, and F. Rendl, *A spectral approach to bandwidth and separator problems in graphs*, manuscript, Institut für Mathematik, Graz, Austria, 1993.
[21] B. Hendrickson and R. Leland, *An improved spectral graph partitioning algorithm for mapping parallel computations*, SIAM J. Sci. Comput., 16 (1995), pp. 452–469.
[22] B. Hendrickson and R. Leland, *Multidimensional Load Balancing*, Tech. report 93-0074, Sandia National Labs, Albuquerque, NM, 1993.
[23] M. Juvan and B. Mohar, *Laplace Eigenvalues and Bandwidth-Type Invariants of Graphs*, Department of Mathematics, University of Ljubljana, Ljubljana, Slovenia, 1990, preprint.
[24] M. Juvan and B. Mohar, *Optimal linear labelings and eigenvalues of graphs*, Discrete Appl. Math., 36 (1992), pp. 153–168.
[25] G. Kumfert and A. Pothen, *A refined spectral algorithm to reduce the envelope and wavefront of sparse matrices*, BIT, to appear, 1997.
[26] J. G. Lewis, *Implementations of the Gibbs–Poole–Stockmeyer and Gibbs–King algorithms*, ACM Trans. Math. Software, 8 (1982), pp. 180–189.
[27] Y. Lin and J. Yuan, *Minimum Profile of Grid Networks in Structure Analysis*, Department of Mathematics, Zhengzhou University, Zhengzhou, People's Republic of China, 1993, preprint.
[28] Y. Lin and J. Yuan, *Profile Minimization Problem for Matrices and Graphs*, Department of Mathematics, Zhengzhou University, Zhengzhou, People's Republic of China, 1993, preprint.
[29] J. W. H. Liu and A. H. Sherman, *Comparative analysis of the Cuthill–Mckee and the reverse Cuthill–Mckee ordering algorithms for sparse matrices*, SIAM J. Numer. Anal., 13 (1976), pp. 198–213.
[30] G. L. Miller, S.-H. Teng, W. Thurston, and S. A. Vavasis, *Automatic mesh partitioning*, in Graph Theory and Sparse Matrix Computation, A. George, J. R. Gilbert, and J. W. H. Liu, eds., IMA Volumes in Mathematics and its Applications, Vol. 56, Springer-Verlag, New York, 1993, pp. 57–84.
[31] B. Mohar and S. Poljak, *Eigenvalues in combinatorial optimization*, in Combinatorial and Graph-Theoretic Problems in Linear Algebra, R. A. Brualdi, ed., Springer-Verlag, New York, 1993, pp. 107–151.
[32] G. H. Paulino, I. F. M. Menezes, M. Gattass, and S. Mukherjee, *Node and element resequencing using the Laplacian of a finite element graph*, Internat. J. Numer. Methods Engrg., 37 (1994), Part I, pp. 1511–1530, Part II, pp. 1531–1555.
[33] 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.
[34] A. Pothen, H. D. Simon, and L. Wang, *Spectral Nested Dissection*, Tech. report CS-92-01, Computer Science, Pennsylvania State University, University Park, PA, 1992.
[35] F. Rendl and H. Wolkowicz, *A projection technique for partitioning the nodes of a graph*, Ann. Oper. Res., 58 (1995), pp. 155–179. (This paper was written in 1990 and appeared in the special issue devoted to the Symposium on Applied Mathematical Programming and Modeling, Budapest, January 1993.)
[36] H. D. Simon, *Partitioning of unstructured problems for parallel processing*, Comput. Systems Engrg., 2 (1991), pp. 135–148.
[37] S. W. Sloan, *An algorithm for profile and wavefront reduction of sparse matrices*, Internat. J. Numer. Methods Engrg., 23 (1986), pp. 239–251.
[38] D. A. Spielman and S.-H. Teng, *Spectral partitioning works: Planar graphs and finite element meshes*, manuscript, Computer Science Department, University of Minnesota, Minneapolis, MN, 1996.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 3, pp. 733–751, July 1997

011

PERTURBATION OF EIGENVALUES OF PRECONDITIONED NAVIER–STOKES OPERATORS*

HOWARD C. ELMAN†

Abstract. We study the sensitivity of algebraic eigenvalue problems associated with matrices arising from linearization and discretization of the steady-state Navier–Stokes equations. In particular, for several choices of preconditioners applied to the system of discrete equations we derive upper bounds on perturbations of eigenvalues as functions of the viscosity and discretization mesh size. The bounds suggest that the sensitivity of the eigenvalues is at worst linear in the inverse of the viscosity and quadratic in the inverse of the mesh size and that scaling can be used to decrease the sensitivity in some cases. Experimental results supplement these results and confirm the relatively mild dependence on viscosity. They also indicate a dependence on the mesh size of magnitude smaller than the analysis suggests.

Key words. eigenvalues, perturbation analysis, Navier–Stokes, preconditioning

AMS subject classifications. Primary, 65F10, 65N20; Secondary, 15A06

PII. S0895479895294873

1. Introduction.

This paper concerns properties of the eigenvalues of matrices arising from the discrete linearized steady-state Navier–Stokes equations. The continuous problem is

$$-\nu\Delta\mathbf{u} + (\mathbf{u}\cdot\operatorname{grad})\,\mathbf{u} + \operatorname{grad} p = \mathbf{f} \quad \text{in } \Omega, \tag{1}$$

together with the incompressibility constraint

$$-\operatorname{div}\mathbf{u} = 0 \quad \text{in } \Omega, \tag{2}$$

subject to suitable boundary conditions on $\partial\Omega$, where Ω is an open bounded domain in $\mathbf{R}^2$. These equations constitute a fundamental problem in computational fluid dynamics; see, e.g., [1], [6], [8]. The two-dimensional vector field $\mathbf{u}$ represents the velocity in Ω, p represents pressure, and the scalar ν is the viscosity, roughly speaking, the ratio of convection to diffusion in the system.

A methodology for computing the numerical solution is to discretize (1)–(2) using finite difference or finite element methods and then to solve the resulting nonlinear system by some iterative method. Linearization leads to a set of matrix equations of the form

$$\begin{pmatrix} F & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ p \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ 0 \end{pmatrix}, \tag{3}$$

where $\mathbf{u}$ and p now represent discrete versions of velocity and pressure, respectively.

We will restrict our attention to the discrete *Oseen equations*

$$\begin{aligned} -\nu\Delta\mathbf{u} + (\mathbf{w}\cdot\operatorname{grad})\,\mathbf{u} + \operatorname{grad} p &= \mathbf{f}, \\ -\operatorname{div}\mathbf{u} &= 0, \end{aligned} \tag{4}$$

* Received by the editors November 17, 1995; accepted for publication (in revised form) by H. Weinberger July 29, 1996. This work was supported by U. S. Army Research Office grant DAAL-0392-G-0016 and U. S. National Science Foundation grants ASC-8958544 and DMS-9423133.
http://www.siam.org/journals/simax/18-3/29487.html
† Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 (elman@cs.umd.edu).

where $\mathbf{w}$ is given such that $\operatorname{div} \mathbf{w} = 0$. These equations arise from a nonlinear iteration of the form $-\nu\Delta\mathbf{u}^{(m)} + (\mathbf{u}^{(m-1)} \cdot \operatorname{grad})\,\mathbf{u}^{(m)} + \operatorname{grad} p^{(m)} = \mathbf{f},\ -\operatorname{div} \mathbf{u}^{(m)} = 0$; see [10]. In this case

$$F = \nu A + N,$$

where A consists of a pair of uncoupled discrete Laplace operators, corresponding to diffusion, and N is a skew-symmetric matrix representing convection. We will also assume that the velocity and pressure discretizations are *div-stable*; see, e.g., [1, p. 57], [8, p. 10ff], [16]. In matrix notation, this is equivalent to the condition

$$\gamma^2 \leq \frac{(p, BA^{-1}B^T p)}{(p, Mp)} \leq \Gamma^2 \quad \text{for all } p, \tag{5}$$

where $(\cdot, \cdot)$ denotes the Euclidean inner product, γ and Γ are constants that are independent of the discretization mesh size h, and for finite elements M is the pressure mass matrix, i.e., the Grammian matrix of basis functions defining the discrete pressure space.[1] For finite differences on uniform grids, a natural analogue is $M = h^2 I$.

Let $\mathcal{L}$ denote the coefficient matrix of (3). The following preconditioning matrices were introduced in [3]: a *block diagonal* preconditioner

$$\mathcal{Q}_D = \begin{pmatrix} F & 0 \\ 0 & \frac{1}{\nu} M \end{pmatrix} \tag{6}$$

and a *block triangular* preconditioner

$$\mathcal{Q}_T = \begin{pmatrix} F & B^T \\ 0 & -\frac{1}{\nu} M \end{pmatrix}. \tag{7}$$

It was shown in [3] that the eigenvalues of each of the preconditioned matrices $\mathcal{A}_D = \mathcal{L}\mathcal{Q}_D^{-1}$ and $\mathcal{A}_T = \mathcal{L}\mathcal{Q}_T^{-1}$ are uniformly bounded independent of the mesh size used in the discretization. Numerical experiments also suggested that Krylov subspace iterative methods such as the generalized minimal residual (GMRES) [13] and quasi-minimal residual (QMR) methods [5] can be used to solve the preconditioned system with iteration counts independent of the mesh size.

We are concerned with the sensitivity of the eigenvalues of the preconditioned Oseen matrix using the two preconditioners (6) and (7). The motivation for studying this lies in the fact that the use of either preconditioner in an iteration entails applying the action of the inverse of the matrix of either (6) or (7) to a vector at each step. This in turn requires the computation of the action of F^{-1}, which, if direct methods are used, will dominate the cost. An alternative that was considered in [3] is to approximate the action of F^{-1} (i.e., compute an approximate solution to systems with coefficient matrix F) using an inner iteration. Unless very stringent stopping criteria are used here, the resulting preconditioned operators can be viewed as perturbations of those of (6)–(7). Thus, we are interested in the sensitivity of the eigenvalues to perturbation.

If the preconditioned matrix is perturbed by a matrix of size ϵ, then the perturbations of the eigenvalues will depend on ϵ and also on parameters associated with the

[1] An inequality analogous to (5) also holds, with different constants, if M is any matrix spectrally equivalent to the mass matrix; for example, M could be the diagonal matrix consisting of the diagonal of the mass matrix [17]. In what follows, we will not distinguish among such possibilities for M.

underlying problem, specifically, the viscosity ν and mesh size h. In this paper, we examine these dependencies using a combination of analytic and experimental results. The analysis derives from Wilkinson's classical perturbation analysis [18], which shows that if there are no nonlinear elementary divisors, the perturbations are of magnitude $O(\epsilon)$. The analytic bounds are stated as functions of ν and h but they also depend on properties of certain matrices associated with the Schur complement $BF^{-1}B^T$ derived from (4). The latter quantities are studied in a series of numerical experiments. The combination of analytic and experimental results indicates that there is an increase in sensitivity to perturbation as the viscosity decreases, with growth roughly linear in $1/\nu$. This effect can be mitigated to some extent by scaling the first equation of (4) (the momentum equation). The bounds also establish linear dependence on $1/h$ with the preconditioner $\mathcal{Q}_T$ and quadratic dependence with $\mathcal{Q}_D$, although the experimental results suggest that perturbations are considerably less sensitive to this parameter.

An outline of the paper now follows. In section 2, we derive preliminary bounds and relations for several operators associated with the preconditioned matrices. In section 3, we derive the analytic perturbation bounds for the block tridiagonal preconditioner, and in section 4, we present the analysis for the block diagonal preconditioner. For simplicity, the analysis is done for the case in which the coefficient matrix of (4) has full rank, although often in practice (and in our experiments) it is rank deficient by one because the pressure p is uniquely defined only up to a constant. In section 5, we show that the analytic results carry over to this case. In section 6, we present the experimental results, and in section 7, we show how the analysis applies for the case of the inexact computation of the action of F^{-1}.

2. Preliminary results. In this section we derive preliminary bounds and relations for several operators associated with the preconditioners (6)–(7). We will assume that the discrete problem (3) arises from a standard finite difference or low-order finite element scheme on a uniform grid with mesh size h and that the discrete problem is scaled so that the extreme eigenvalues of the discrete Laplace operators of A are contained in an interval of the form $[c_1h^2, c_2]$, where here and below c_i denotes a generic constant that is independent of h and ν. This is a natural scaling for finite elements, and for finite differences on a uniform grid it corresponds to the five-point operator with 4 in the diagonal entries and -1 in the off-diagonal entries. With this normalization, BB^T/h^2 is also a scaled discrete Laplace operator and its eigenvalues are contained in an interval of the same form. Let the discrete velocity and pressure spaces have dimension n_u and n_p, respectively. For div-stable discretizations, $n_u \geq n_p$, and typically n_u is significantly larger than n_p.[2]

It will be convenient to use the symbol Q instead of $\frac{1}{\nu}M$ in the matrices of (6). The preconditioned matrices are then given by

$$\mathcal{A}_D = \begin{pmatrix} F & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} F^{-1} & 0 \\ 0 & Q^{-1} \end{pmatrix} = \begin{pmatrix} I & K^T \\ G & 0 \end{pmatrix} \tag{8}$$

for the block diagonal preconditioner and

$$\mathcal{A}_T = \begin{pmatrix} F & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} F^{-1} & F^{-1}B^TQ^{-1} \\ 0 & -Q^{-1} \end{pmatrix} = \begin{pmatrix} I & 0 \\ G & H \end{pmatrix} \tag{9}$$

for the block triangular preconditioner. The submatrices on the right of (8)–(9) are

$$G = BF^{-1}, \quad K^T = B^TQ^{-1}, \quad H = GK^T = BF^{-1}B^TQ^{-1}.$$

[2] For two-dimensional problems, the vector $\mathbf{u}$ has two components of grid vectors, and stability considerations often also lead to more grid points for velocity than for pressure [8].

The identity matrices are of order n_u.

We recall some results from [3], which give bounds on the eigenvalues of H. Let $S = BF^{-1}B^T$ denote the Schur complement matrix for (3), and let

$$C = B\left(\frac{F^{-1}+F^{-T}}{2}\right)B^T, \qquad R = B\left(\frac{F^{-1}-F^{-T}}{2}\right)B^T$$

denote the symmetric and skew-symmetric parts of S, respectively. It was shown in [3] that

$$\frac{\gamma^2\nu^2}{\delta^2+\nu^2} \le \frac{(q,Cq)}{(q,Qq)} \le \Gamma^2, \qquad \frac{|(q,Rq)|}{(q,Qq)} \le \frac{\Gamma^2}{2}, \tag{10}$$

where γ and Γ are as in (5) and δ is the largest eigenvalue of $A^{-1}N$, which is also uniformly bounded independent of h [4]. Consequently, Bendixson's theorem [14, p. 418] implies that the eigenvalues of H are contained in the box

$$\left[\frac{\gamma^2\nu^2}{\delta^2+\nu^2}, \Gamma^2\right] \times \left[\frac{-\Gamma^2}{2}, \frac{\Gamma^2}{2}\right] \tag{11}$$

in the complex plane.

We first derive bounds on the singular values of G, which will be used in the perturbation analysis for the block triangular preconditioner.

THEOREM 2.1. *The largest singular value of G is bounded above by a quantity of magnitude $O(1/\nu)$ which is independent of h as $h \to 0$. The smallest singular value of G is bounded below by a quantify of magnitude $O(h)$ which is bounded independent of ν as $\nu \to 0$.*

Proof. The singular values of G are the square roots of the eigenvalues of GG^T, and the largest and smallest of these eigenvalues are the extrema of $(q, GG^Tq)/(q,q)$. This Rayleigh quotient can be rewritten as

$$\frac{(q,GG^Tq)}{(q,q)} = \frac{(F^{-T}B^Tq, F^{-T}B^Tq)}{(q,q)} = \frac{(F^{-T}B^Tq, F^{-T}B^Tq)}{(F^{-T}B^Tq, B^Tq)}\frac{(F^{-T}B^Tq, B^Tq)}{(q,q)}. \tag{12}$$

We consider the two terms in the product on the right of (12) separately. For the first term, the substitution $w = F^{-T}B^Tq$ gives

$$\frac{(F^{-T}B^Tq, F^{-T}B^Tq)}{(F^{-T}B^Tq, B^Tq)} = \frac{(w,w)}{(w,F^Tw)} = \frac{(w,w)}{\left(w, \left(\frac{F+F^T}{2}\right)w\right)} = \frac{1}{\nu}\frac{(w,w)}{(w,Aw)}.$$

Under the assumption on the scaling of the discrete Laplacian operators composing A, it follows that

$$\frac{c_1}{\nu} \le \frac{(F^{-T}B^Tq, F^{-T}B^Tq)}{(F^{-T}B^Tq, B^Tq)} \le \frac{c_2h^{-2}}{\nu}. \tag{13}$$

The second term of the product in (12) is

$$\frac{(q,BF^{-1}B^Tq)}{(q,q)} = \frac{(q,Cq)}{(q,q)} = \frac{(q,Cq)}{(q,Qq)}\frac{(q,Qq)}{(q,q)}. \tag{14}$$

It is well known (see [17]) that the pressure mass matrix is spectrally equivalent to $h^2 I$, so that

$$\frac{c_1 h^2}{\nu} \leq \frac{(q, Qq)}{(q, q)} \leq \frac{c_2 h^2}{\nu}\,. \tag{15}$$

Thus, the bounds for the symmetric part in (10) together with (12)–(15) imply

$$\frac{c_1 h^2}{\delta^2 + \nu^2} \leq \frac{(q, GG^T q)}{(q, q)} \leq \frac{c_2}{\nu^2}\,. \qquad \square$$

The singular values of K will be used to analyze the block diagonal preconditioner.

LEMMA 2.2. *The largest singular value of K is bounded above by a quantity of magnitude $O\left(\frac{\nu}{h}\right)$. The smallest singular value is bounded below by a quantity of magnitude $O(\nu)$.*

Proof. The largest singular value is $\|K\|_2 = \|K^T\|_2$. Using $K^T = B^T Q^{-1}$ and $Q = \frac{1}{\nu} M$, we have

$$\|K^T\|_2 \leq \nu \, \|B^T\|_2 \, \|M^{-1}\|_2.$$

But $\|M^{-1}\|_2 = O(h^{-2})$, and our assumptions on B imply that $\|B^T\|_2 = \|BB^T\|_2^{1/2} = O(h)$. The smallest singular value is the inverse of $\|(KK^T)^{-1}\|_2^{1/2}$. Then

$$\|(KK^T)^{-1}\|_2 \leq \frac{1}{\nu^2} \, \|M\|_2^2 \, \|(BB^T)^{-1}\|_2 = \frac{1}{\nu^2} \, c_1 h^4 \, c_2 h^{-4}. \qquad \square$$

Consider an alternative scaling in problems (1) and (4) in which the first equation is multiplied by $\frac{1}{\nu}$. This does not change the solutions, but, as we will show in sections 3 and 4, it affects the sensitivity of discrete eigenvalues. For (4), scaling gives

$$-\Delta \mathbf{u} + \frac{1}{\nu} (\mathbf{w} \cdot \operatorname{grad}) \, \mathbf{u} + \operatorname{grad} \left(\frac{1}{\nu} p \right) = \frac{1}{\nu} \mathbf{f}. \tag{16}$$

The new discrete problem is as in (3) except that F, p, and $\mathbf{f}$ are replaced by $\hat{F} = \frac{1}{\nu} F$, $\frac{1}{\nu} p$, and $\frac{1}{\nu} \mathbf{f}$, respectively. Let $\hat{Q} = \nu Q = M$, $\hat{G} = B\hat{F}^{-1}$, $\hat{K}^T = B^T \hat{Q}^{-1}$, and $\hat{H} = \hat{G}\hat{K}^T$. The analogues of (11) and the bounds of Theorem 2.1 and Lemma 2.2 are given below. The proof follows from the fact that $\hat{G} = \nu G$, $\hat{K} = \frac{1}{\nu} K$, and $\hat{H} = H$.

COROLLARY 2.3. *With the scaling of* (16)*, the eigenvalues of $\hat{H}$ are contained in the box* (11)*; the singular values of $\hat{G}$ are bounded above by a quantity of magnitude $O(1)$ and below by a quantity of magnitude $O(h\nu)$; and the singular values of $\hat{K}$ are bounded above by a quantity of magnitude $O\left(\frac{1}{h}\right)$ and below by a quantity of magnitude $O(1)$.*

In the following, we will not specifically identify the matrices associated with this scaling using the "hat" symbol. Instead, we will use the notation of (8)–(9) to refer generically to both scalings. We will distinguish them as derived from either the "original" formulation (4) or the "scaled" formulation (16) of the Oseen equations.

Finally, we will use the notation $\alpha_{G,K}$ to denote the secant of the largest principal angle between Range(G^T) and Range(K^T). That is, if Q_G and Q_K are matrices whose columns represent orthonormal bases of Range(G^T) and Range(K^T), respectively, then

$$\alpha_{G,K} = \|(Q_G^T Q_K)^{-1}\|_2 = \frac{1}{\sigma_{min}(Q_G^T Q_K)}\,, \tag{17}$$

where σ_{min} denotes the smallest singular value (see [7, p. 584]). It is easily shown (e.g., using QR decompositions) that

$$\alpha_{G,K} = \|K^T(GK^T)^{-1}G\|_2. \tag{18}$$

3. Analytic bounds for the block triangular preconditioner. It is evident from (9) that the eigenvalues of $\mathcal{A}_T$ consist of $\lambda = 1$ of multiplicity n_u together with the eigenvalues of H. We seek a factorization

$$\mathcal{A}_T = \mathcal{V}_T \mathcal{D}_T \mathcal{V}_T^{-1} \tag{19}$$

that provides insight into the sensitivity of these eigenvalues to perturbation. We will look for factors of the form

$$\mathcal{D}_T = \begin{pmatrix} I & 0 \\ D_{21} & \Lambda \end{pmatrix} \begin{matrix} n_u \\ n_p \end{matrix}, \qquad \mathcal{V}_T = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \begin{matrix} n_u \\ n_p \end{matrix}, \tag{20}$$
$$\begin{matrix} n_u & n_p \end{matrix} \qquad\qquad \begin{matrix} n_u & n_p \end{matrix}$$

where the subblocks of $\mathcal{D}_T$ and $\mathcal{V}_T$ must be determined and the dimensions are as indicated. Let H have Jordan canonical form $H = P\Lambda P^{-1}$. The requirement $\mathcal{A}_T\mathcal{V}_T = \mathcal{V}_T\mathcal{D}_T$ is satisfied if

$$\begin{aligned} &V_{11} = I, && V_{12} = 0, \\ &G = (I - H)V_{21} + PD_{21}, && V_{22} = P. \end{aligned} \tag{21}$$

We distinguish between two cases: $1 \in \sigma(H)$ and $1 \notin \sigma(H)$.

Suppose first that H has no eigenvalues equal to 1. The choice $D_{21} = 0$ in (20) leads to $V_{21} = (I - H)^{-1}G$. In this case, (19) represents a Jordan form for $\mathcal{A}_T$ with

$$\mathcal{V}_T = \begin{pmatrix} I & 0 \\ (I - H)^{-1}G & P \end{pmatrix}, \qquad \mathcal{V}_T^{-1} = \begin{pmatrix} I & 0 \\ -P^{-1}(I - H)^{-1}G & P^{-1} \end{pmatrix}. \tag{22}$$

Let $\mathcal{A}_T(\epsilon) = \mathcal{A}_T + \epsilon\mathcal{E}$ be a perturbation of $\mathcal{A}_T$, where

$$\mathcal{E} = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}. \tag{23}$$

The classical perturbation analysis of Wilkinson based on Gerschgorin theory [18, p. 71ff] shows that for every eigenvalue λ of $\mathcal{A}_T$ with only linear elementary divisors, perturbations of λ in $\sigma(\mathcal{A}_T(\epsilon))$ are contained in a circle centered at λ with radius of size $c\epsilon$, where c is independent of ϵ. The structure of $\mathcal{V}_T$ can be used to obtain further insight into the sizes of the perturbations. Let

$$\hat{\mathcal{E}}_T = \mathcal{V}_T^{-1}\mathcal{E}\mathcal{V}_T = \begin{pmatrix} \hat{E}_{11} & \hat{E}_{12} \\ \hat{E}_{21} & \hat{E}_{22} \end{pmatrix}, \tag{24}$$

so that we are concerned with the eigenvalues of

$$\mathcal{V}_T^{-1}\mathcal{A}_T(\epsilon)\mathcal{V}_T = \mathcal{D}_T + \epsilon\hat{\mathcal{E}}_T = \begin{pmatrix} I & 0 \\ 0 & \Lambda \end{pmatrix} + \epsilon \begin{pmatrix} \hat{E}_{11} & \hat{E}_{12} \\ \hat{E}_{21} & \hat{E}_{22} \end{pmatrix}. \tag{25}$$

Here and in the following the symbol "c" represents a generic constant that is independent of the parameters h, ν, and ϵ.

LEMMA 3.1. *If $1 \notin \sigma(H)$, then there are n_u eigenvalues $\hat{\lambda}$ of $\mathcal{A}_T(\epsilon)$ (counting multiplicity) satisfying*

$$|\hat{\lambda} - 1| \le \epsilon \, \|\hat{E}_{11}\|_\infty + c\epsilon^2 \, \|\hat{E}_{12}\|_\infty. \tag{26}$$

If there are $m \le n_p$ eigenvalues λ of H with linear elementary divisors, then there are m eigenvalues $\hat{\lambda}$ of $\mathcal{D}_T(\epsilon)$, distinct from those of (26)*, that satisfy*

$$|\hat{\lambda} - \lambda| \le \epsilon \, \|\hat{E}_{22}\|_\infty + c\epsilon^2 \, \|\hat{E}_{21}\|_\infty. \tag{27}$$

Proof. Multiplying the first block row on the right side of (25) by ϵ/β, multiplying the first block column by β/ϵ, and then applying Gerschgorin's theorem leads to the bound

$$|\hat{\lambda} - (1 + \epsilon[\hat{E}_{11}]_{ii})| \le \epsilon \sum_{j \ne i} |[\hat{E}_{11}]_{ij}| + \epsilon^2/\beta \sum_j |[\hat{E}_{12}]_{ij}|,$$

where β is such that the Gerschgorin disks for the first and second blocks of the scaled matrix are disjoint (see [18, p. 73]). Assertion (26) follows. The argument for (27) is identical, applied to the second block row of (25). ∎

Thus, we can restrict our attention to the block diagonal entries of $\hat{\mathcal{E}}_T$. Using (22)–(24) we have

$$\begin{aligned} \hat{E}_{11} &= E_{11} + E_{12}(I - H)^{-1}G\,, \\ \hat{E}_{22} &= -P^{-1}(I - H)^{-1}GE_{12}P + P^{-1}E_{22}P. \end{aligned} \tag{28}$$

The following result gives bounds on the perturbations of eigenvalues as functions of the viscosity ν and mesh size h for a perturbation satisfying $\|\mathcal{E}\|_2 \le 1$.[3]

THEOREM 3.2. *Assume $\|\mathcal{E}\|_2 \le 1$. If $1 \notin \sigma(H)$, then the eigenvalues $\hat{\lambda}$ of $\mathcal{A}_T(\epsilon)$ that are perturbations of $\lambda = 1$ satisfy*

$$|\hat{\lambda} - 1| \le \begin{cases} \epsilon \, \dfrac{c_1}{\nu h} + O(\epsilon^2) & \textit{for the original formulation}, \\ \epsilon \, \dfrac{c_1}{h} + O(\epsilon^2) & \textit{for the scaled formulation}, \end{cases} \tag{29}$$

where $c_1 = c \, \|(I - H)^{-1}\|_2$. For eigenvalues λ of H with linear elementary divisors, the perturbations $\hat{\lambda} \in \sigma(\mathcal{A}_T(\epsilon))$ satisfy

$$|\hat{\lambda} - \lambda| \le \begin{cases} \epsilon \, \dfrac{c_2}{\nu h} + O(\epsilon^2) & \textit{for the original formulation}, \\ \epsilon \, \dfrac{c_2}{h} + O(\epsilon^2) & \textit{for the scaled formulation}, \end{cases} \tag{30}$$

where $c_2 = c \, \kappa(P) \, \|(I - H)^{-1}\|_2$.

Proof. Relations (28) together with standard bounds on matrix l_p-norms give

$$\begin{aligned} \|\hat{E}_{11}\|_\infty &\le \frac{c}{h} \, \|\hat{E}_{11}\|_2 \le \frac{c}{h} \left(1 + \|(I - H)^{-1}\|_2 \, \|G\|_2\right), \\ \|\hat{E}_{22}\|_\infty &\le \frac{c}{h} \, \|\hat{E}_{22}\|_2 \le \frac{c}{h} \, \kappa(P) \left(\|(I - H)^{-1}\|_2 \, \|G\|_2 + 1\right). \end{aligned}$$

[3] This assumption is stronger than the inequality $|\mathcal{E}_{ij}| \le 1$ used by Wilkinson; the latter condition follows from our assumption.

The conclusions follow from the upper bounds on $\|G\|_2$ in Theorem 2.1. ∎

In the case $1 \in \sigma(H)$, we use an alternative version of (20). By analogy with the analysis above, let $V_{21} = (I-H)^\dagger G$, where $(I-H)^\dagger$ is the pseudoinverse. Then

$$\mathcal{V}_T = \begin{pmatrix} I & 0 \\ (I-H)^\dagger G & P \end{pmatrix}, \qquad \mathcal{V}_T^{-1} = \begin{pmatrix} I & 0 \\ -P^{-1}(I-H)^\dagger G & P^{-1} \end{pmatrix}.$$

Similarity (19) then holds for the choice

$$\mathcal{D}_T = \begin{pmatrix} I & 0 \\ P^{-1}\Pi G & \Lambda \end{pmatrix},$$

where Π is the orthogonal projection onto the null space of $I - H^T$. The perturbation of $\mathcal{D}_T$ associated with $\mathcal{A}_T(\epsilon)$ is then

$$\mathcal{V}_T^{-1}\mathcal{A}_T(\epsilon)\mathcal{V}_T = \mathcal{D}_T + \epsilon\hat{\mathcal{E}}_T = \begin{pmatrix} I & 0 \\ 0 & \Lambda \end{pmatrix} + \begin{pmatrix} \epsilon\hat{E}_{11} & \epsilon\hat{E}_{12} \\ -P^{-1}\Pi G + \epsilon\hat{E}_{21} & \epsilon\hat{E}_{22} \end{pmatrix}. \tag{31}$$

First consider the eigenvalues of H different from 1 with linear elementary divisors. Let η be a parameter in $(0,1)$. Multiplying the second block row of (31) by ϵ^η and the second block column by $\epsilon^{-\eta}$ produces the matrix

$$\begin{pmatrix} I & 0 \\ 0 & \Lambda \end{pmatrix} + \begin{pmatrix} \epsilon\hat{E}_{11} & \epsilon^{1-\eta}\hat{E}_{12} \\ -\epsilon^\eta P^{-1}\Pi G + \epsilon^{1+\eta}\hat{E}_{21} & \epsilon\hat{E}_{22} \end{pmatrix}. \tag{32}$$

For any $\eta \in (0,1)$ and small enough ϵ, the Gerschgorin disk for a perturbation $\hat{\lambda}$ of $\lambda \in \sigma(H)$, $\lambda \neq 1$, is disjoint from the disks corresponding to perturbations of the eigenvalue 1. Consequently, if λ has only linear elementary divisors, then

$$|\hat{\lambda} - \lambda| \leq \epsilon^\eta \, \|P^{-1}\Pi G\|_\infty + O(\epsilon).$$

For eigenvalues equal to 1, multiplying the second block row of (31) by $\epsilon^{1/2}$ and multiplying the second block column by $\epsilon^{-1/2}$ produces

$$\begin{pmatrix} I & 0 \\ 0 & \Lambda \end{pmatrix} + \begin{pmatrix} \epsilon\hat{E}_{11} & \epsilon^{1/2}\hat{E}_{12} \\ -\epsilon^{1/2}P^{-1}\Pi G + \epsilon^{3/2}\hat{E}_{21} & \epsilon\hat{E}_{22} \end{pmatrix}. \tag{33}$$

Thus, for $\lambda = 1$ with linear or quadratic elementary divisors, the perturbations satisfy

$$|\hat{\lambda} - \lambda| \leq \epsilon^{1/2} \, \max\left(\|\hat{E}_{12}\|_\infty, \|P^{-1}\Pi G\|_\infty\right) + O(\epsilon).$$

Bounding the matrix infinity norms in (32) and (33) gives the following result.

THEOREM 3.3. *If $1 \in \sigma(H)$ with linear or quadratic elementary divisors, then the eigenvalues $\hat{\lambda}$ of $\mathcal{A}_T(\epsilon)$ that are perturbations of $\lambda = 1$ satisfy*

$$|\hat{\lambda} - 1| \leq \epsilon^{1/2} \, \frac{c}{h} \, \max\left(\|P\|_2, \|P^{-1}\|_2 \, \|G\|_2\right) + O(\epsilon). \tag{34}$$

For $\lambda \in \sigma(H)$ different from 1 with linear elementary divisors, the perturbations $\hat{\lambda} \in \sigma(\mathcal{A}_T(\epsilon))$ satisfy

$$|\hat{\lambda} - \lambda| \leq \epsilon^\eta \, \frac{c}{h} \|P^{-1}\|_2 \|G\|_2 + O(\epsilon) \tag{35}$$

for any $\eta \in (0,1)$. Here $\|G\|_2 = O(1/\nu)$ in the original formulation and $\|G\|_2 = O(1)$ for the scaled formulation.

Note that for any $\lambda \in \sigma(\mathcal{A}_T)$ with nonlinear elementary divisors, bounds analogous to (29)–(30) and (34)–(35) can be obtained in which the dependence on ϵ is of the form $\epsilon^{1/r}$, where r is the order of the largest Jordan block for λ; see [18, p. 79].

4. Analytic bounds for the block diagonal preconditioner. For the block diagonal preconditioner, we have results only in the case in which the Schur complement matrix H has no nonlinear elementary divisors. As in Theorem 3.2, the perturbation analysis will be stated in terms of properties of P, a fixed matrix of eigenvectors of H. In addition, some of the bounds depend on $\alpha_{G,K}$, the secant of the largest principal angle between Range(G^T) and Range(K^T), as defined in (17).

Consider the eigenvalue problem for $\mathcal{A}_D$,

$$\begin{pmatrix} I & K^T \\ G & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \theta \begin{pmatrix} u \\ p \end{pmatrix}.$$

This leads to the conditions

$$K^T p = (\theta - 1)u, \qquad Gu = \theta p \tag{36}$$

on the components of the eigenvectors. One solution corresponds to the eigenvalue $\theta = 1$; the assumption that B and therefore K have full rank then implies that any associated eigenvector satisfies $p = 0$, $Gu = 0$. We use two approaches to identify eigenvalues $\theta \neq 1$ and corresponding eigenvectors.

1. The substitution of $u = (\frac{1}{\theta - 1})K^T p$ into the second equation of (36) gives

$$GK^T p = \theta(\theta - 1)\, p.$$

That is, any eigenpair (λ, p) of $H = GK^T$ leads to two eigenvalues

$$\theta_+ = \frac{1 + \sqrt{1 + 4\lambda}}{2}, \qquad \theta_- = \frac{1 - \sqrt{1 + 4\lambda}}{2} \tag{37}$$

of $\mathcal{A}_D$. The associated eigenvectors are

$$\begin{pmatrix} u_+ \\ p \end{pmatrix}, \qquad \begin{pmatrix} u_- \\ p \end{pmatrix}, \tag{38}$$

where

$$u_+ = \left(\frac{1}{\theta_+ - 1}\right) K^T p, \qquad u_- = \left(\frac{1}{\theta_- - 1}\right) K^T p. \tag{39}$$

2. Alternatively, the substitution of $p = \frac{1}{\theta} Gu$ into the first equation of (36) gives

$$K^T Gu = \theta(\theta - 1)\, u.$$

$K^T G$ has a zero eigenvalue of multiplicity $n_u - n_p$ (the dimension of the null space of G) plus $2n_p$ nonzero eigenvalues. If (λ, u) is an eigenpair with $\lambda \neq 0$, then (37) defines a pair of eigenvalues of $\mathcal{A}_D$. In this case, the eigenvectors have the form

$$\begin{pmatrix} u \\ p_+ \end{pmatrix}, \qquad \begin{pmatrix} u \\ p_- \end{pmatrix}, \tag{40}$$

where

$$p_+ = \left(\frac{1}{\theta_+}\right) Gu, \qquad p_- = \left(\frac{1}{\theta_-}\right) Gu. \tag{41}$$

It is straightforward to show that any u_+ or u_- from (39) is an eigenvector of K^TG and any p_+ or p_- from (41) is an eigenvector of GK^T. Therefore, we will use the symbol "$\pm$" to refer to the pairs of eigenvalues and eigenvectors of $\mathcal{A}_D$; i.e., $u_\pm$ will refer to the first entries of either (38) or (40), $p_\pm$ to the second entries, and $\theta_\pm$ to the associated eigenvalue. We have established the following result.

THEOREM 4.1. *The eigenvectors of $\mathcal{A}_D$ corresponding to the eigenvalue $\theta = 1$ have the form $\begin{pmatrix} u \\ 0 \end{pmatrix}$, where $Gu = 0$. The eigenvectors corresponding to eigenvalues different from 1 have the form $\begin{pmatrix} u_\pm \\ p_\pm \end{pmatrix}$, where the associated eigenvalues $\theta_\pm$ satisfy* (37), *and*

$$Hp_\pm = \lambda p_\pm, \quad K^TGu_\pm = \lambda u_\pm, \quad p_\pm = \left(\frac{1}{\theta_\pm}\right) Gu_\pm, \quad u_\pm = \left(\frac{1}{\theta_\pm - 1}\right) K^T p_\pm.$$

Let $\Theta_\pm$ denote a diagonal matrix with entries $\{\theta_\pm\}$ from (37). *If the eigenvalues of H have only linear elementary divisors, then $\mathcal{A}_D\mathcal{V}_D = \mathcal{V}_D\mathcal{D}_D$, where*

$$\mathcal{D}_D = \begin{pmatrix} I & 0 & 0 \\ 0 & \Theta_+ & 0 \\ 0 & 0 & \Theta_- \end{pmatrix} \begin{matrix} n_u - n_p \\ n_p \\ n_p \end{matrix}, \qquad \mathcal{V}_D = \begin{pmatrix} V & U_+ & U_- \\ & & \\ 0 & P_+ & P_- \end{pmatrix} \begin{matrix} n_u \\ \\ n_p \end{matrix},$$
$$\begin{matrix} {\scriptstyle n_u -} \\ {\scriptstyle n_p} \end{matrix} \; n_p \; n_p \qquad\qquad \begin{matrix} {\scriptstyle n_u -} \\ {\scriptstyle n_p} \end{matrix} \; n_p \, n_p$$

in which the columns of V form an orthogonal basis of the null space of G and

$$P_\pm = GU_\pm\Theta_\pm^{-1}, \quad U_\pm = K^TP_\pm(\Theta_\pm - I)^{-1}. \tag{42}$$

We will normalize the matrices $U_\pm$ and $P_\pm$ as follows:

$$\begin{aligned} U_+ &= U_- = U = K^TP, \\ P_+ &= P(\Theta_+ - I), \\ P_- &= P(\Theta_- - I). \end{aligned} \tag{43}$$

It is then easily verified that

$$\mathcal{V}_D^{-1} = \begin{pmatrix} I & 0 & 0 \\ 0 & (\Theta_+ - \Theta_-)^{-1} & 0 \\ 0 & 0 & (\Theta_- - \Theta_+)^{-1} \end{pmatrix} \begin{pmatrix} \Phi & 0 \\ P_+^{-1}G & P^{-1} \\ P_-^{-1}G & P^{-1} \end{pmatrix},$$

where

$$\Phi = V^T(I - U(GU)^{-1}G) = V^T(I - K^T(GK^T)^{-1}G). \tag{44}$$

Now let $\mathcal{A}_D(\epsilon) = \mathcal{A}_D + \epsilon\mathcal{E}$ be a perturbation of $\mathcal{A}_D$, where $\mathcal{E}$ is as in (23). The eigenvalues of this perturbed matrix are the same as those of

$$\mathcal{V}_D^{-1}\mathcal{A}_D(\epsilon)\mathcal{V}_D = \mathcal{D}_D + \epsilon\hat{\mathcal{E}}_D = \begin{pmatrix} I + \epsilon\hat{E}_{11} & \epsilon\hat{E}_{12} & \epsilon\hat{E}_{13} \\ \epsilon\hat{E}_{21} & \Theta_+ + \epsilon\hat{E}_{22} & \epsilon\hat{E}_{23} \\ \epsilon\hat{E}_{31} & \epsilon\hat{E}_{32} & \Theta_- + \epsilon\hat{E}_{33} \end{pmatrix}, \tag{45}$$

where $\hat{\mathcal{E}}_D = \mathcal{V}_D^{-1}\mathcal{E}\mathcal{V}_D$. As in section 3, Gerschgorin analysis implies that for small ϵ the effects of perturbation can be bounded using the perturbations on the block

diagonal of (45). These are given by

$$
\begin{aligned}
\hat{E}_{11} &= \Phi^T E_{11} V, \\
\hat{E}_{22} &= (\Theta_+ - \Theta_-)^{-1} \left[P_+^{-1} G (E_{11} U + E_{12} P_+) + P^{-1} (E_{21} U + E_{22} P_+) \right], \\
\hat{E}_{33} &= (\Theta_- - \Theta_+)^{-1} \left[P_-^{-1} G (E_{11} U + E_{12} P_-) + P^{-1} (E_{21} U + E_{22} P_-) \right].
\end{aligned} \tag{46}
$$

THEOREM 4.2. *If H has no nonlinear elementary divisors, then the eigenvalues $\hat{\theta}$ of $\mathcal{A}_D(\epsilon)$ that are perturbations of $\theta = 1$ satisfy*

$$
|\hat{\theta} - 1| \le \epsilon \, \frac{c}{h} \, (1 + \alpha_{G,K}) + O(\epsilon^2)
$$

for both the original and scaled formulations of the discrete Oseen equations. The perturbations of eigenvalues different from 1 *satisfy*

$$
|\hat{\theta}_+ - \theta_+| \le \begin{cases} \epsilon \, \kappa(P) \, \alpha_{G,K} \left(\dfrac{c_1}{h^2} + \dfrac{c_2}{\nu h} \right) + O(\epsilon^2) & \text{for the original formulation,} \\ \epsilon \, \kappa(P) \, \alpha_{G,K} \, \dfrac{c}{h^2} + O(\epsilon^2) & \text{for the scaled formulation} \end{cases} \tag{47}
$$

and

$$
|\hat{\theta}_- - \theta_-| \le \begin{cases} \epsilon \, \kappa(P) \left(\dfrac{c_1}{h^2} + \dfrac{c_2}{\nu h} \right) + O(\epsilon^2) & \text{for the original formulation,} \\ \epsilon \, \kappa(P) \, \dfrac{c}{h^2} + O(\epsilon^2) & \text{for the scaled formulation.} \end{cases} \tag{48}
$$

Proof. For the perturbations of $\lambda = 1$, it follows from (18), (44), and (45) that

$$
|\hat{\theta} - 1| \le \epsilon \, c \, \|\hat{E}_{11}\|_\infty \le \epsilon \, \frac{c}{h} \, \|\hat{E}_{11}\|_2 \le \epsilon \, \frac{c}{h} \, (1 + \alpha_{G,K}).
$$

For the perturbations of eigenvalues different from 1, note that (37) and the bounds (11) on the eigenvalues of H imply that the norm of each of the diagonal matrices

$$
\begin{array}{lll} \Theta_+ - I, & \Theta_- - I, & (\Theta_+ - \Theta_-)^{-1}, \\ \Theta_+^{-1}, & (\Theta_- - I)^{-1} & \end{array} \tag{49}
$$

is bounded independent of ν and h. Gerschgorin analysis gives

$$
|\hat{\theta}_\pm - \theta_\pm| \le \epsilon \, \frac{c}{h} \, \|\hat{E}_{jj}\|_2, \tag{50}
$$

where $j = 2$ corresponds to Θ_+ and $j = 3$ to Θ_-. Relations (43) and (46) then imply

$$
\|\hat{E}_{jj}\|_2 \le \|(\Theta_+ - \Theta_-)^{-1}\|_2 \left(\|P_\pm^{-1} G\|_2 + \|P^{-1}\|_2 \right) \left(\|K^T\|_2 + \|\Theta_\pm - I\|_2 \right) \|P\|_2. \tag{51}
$$

Consider $\|P_-^{-1} G\|_2$. Using the expression for P_- in (43), we have

$$
\|P_-^{-1} G\|_2 \le \|(\Theta_- - I)^{-1}\|_2 \, \|P^{-1}\|_2 \, \|G\|_2,
$$

and combining this with (51) gives

$$
\|\hat{E}_{33}\|_2 \le \kappa(P) \, \|(\Theta_- - \Theta_+)^{-1}\|_2 \left(\|(\Theta_- - I)^{-1}\|_2 \, \|G\|_2 + 1 \right) \left(\|K^T\|_2 + \|\Theta_- - I\|_2 \right).
$$

Result (48) follows from (50), the boundedness of the matrices of (49), and the bounds on $\|G\|_2$ and $\|K^T\|_2$ in Lemma 2.2 and Corollary 2.3.

For $\|P_+^{-1}G\|_2$, first observe that (42) and (43) imply

$$(P_+^{-1}G)(K^T P\Theta_+^{-1}) = I.$$

Let $(P_+^{-1}G)^T = Q_1R_1$ and $K^TP\Theta_+^{-1} = Q_2R_2$ be QR factorizations, where R_1 and R_2 are square and nonsingular. Then $R_1^TQ_1^TQ_2R_2 = I$ and

$$\|P_+^{-1}G\|_2 = \|R_1\|_2 \le \|R_2^{-1}\|_2\,\|(Q_1^TQ_2)^{-1}\|_2 = \alpha_{G,K}\,\|R_2^{-1}\|_2. \tag{52}$$

But $\|R_2^{-1}\|_2$ is the inverse of the smallest singular value of $K^TP\Theta_+^{-1}$, so that

$$\|R_2^{-1}\|_2 \le \|\Theta_+P^{-1}(KK^T)^{-1}P^{-T}\Theta_+\|_2^{1/2} \le \|\Theta_+\|_2\,\|P^{-1}\|_2\,\|(KK^T)^{-1}\|_2^{1/2}.$$

From Lemma 2.2, Corollary 2.3, and the boundedness of $\|\Theta_+\|_2$, the term on the right is bounded by $c/\nu\,\|P^{-1}\|_2$ for the original formulation and $c\,\|P^{-1}\|_2$ for the alternative formulation. Result (47) then follows from (50), (51) (with $j = 2$), and (52). ∎

Remark 4.1. The difference between (47) and (48) stems from the fact that $\|(\Theta_+ - I)^{-1}\|_2$ is *not* independent of ν, so we cannot bound $\|P_+^{-1}G\|_2$ directly. The results of section 6 suggest that the perturbations do not behave differently, but we see no way to avoid including $\alpha_{G,K}$ in (47).

5. The rank-deficient case. Unless an additional constraint is imposed on the pressure in (1)–(2) or (4), the matrix of (3) will be rank deficient by one. This is the case for the test problems of section 6. We outline here how the analysis above carries over in the rank-deficient case.

For the block triangular preconditioner, $\mathcal{A}_T\mathcal{V}_T = \mathcal{V}_T\mathcal{D}_T$ where, for $1 \notin \sigma(H)$, the analogues of the matrices defined in (20)–(22) are

$$\mathcal{D}_T = \begin{pmatrix} I & 0 \\ 0 & \Lambda \end{pmatrix},\ \mathcal{V}_T = \begin{pmatrix} I & 0 \\ (I-H)^{-1}G & P \end{pmatrix},\ \mathcal{V}_T^{\dagger} = \begin{pmatrix} I & 0 \\ -P^{\dagger}(I-H)^{-1}G & P^{\dagger} \end{pmatrix}.$$

Λ is a square matrix of order $n_p - 1$ whose eigenvalues are the nonzero eigenvalues of H, and P spans the associated invariant subspace. In particular, $P^{\dagger}$ replaces P^{-1} and $\mathcal{V}_T^{\dagger}$ replaces $\mathcal{V}_T^{-1}$. The analysis of perturbations of the nonzero eigenvalues of $\mathcal{A}_T$ then carries through verbatim with the inverse of the smallest singular value of P in place of $\|P^{-1}\|_2$. The case $1 \in \sigma(H)$ is generalized in a similar manner with $(I-H)^{\dagger}$ in place of $(I-H)^{-1}$.

For the block diagonal preconditioner, the analogue of the similarity transformation of Theorem 4.1 is $\mathcal{A}_D\mathcal{V}_D = \mathcal{V}_D\mathcal{D}_D$, where

$$\mathcal{D}_D = \begin{pmatrix} I & 0 & 0 \\ 0 & \Theta_+ & 0 \\ 0 & 0 & \Theta_- \end{pmatrix} \begin{matrix} n_u - n_p + 1 \\ n_p \\ n_p \end{matrix}, \qquad \mathcal{V}_D = \begin{pmatrix} V & U_+ & U_- \\ & & \\ 0 & P_+ & P_- \end{pmatrix} \begin{matrix} n_u \\ \\ n_p \end{matrix}.$$

$$\begin{matrix} n_u - \\ n_p + 1 \end{matrix} \quad \begin{matrix} n_p \\ -1 \end{matrix} \quad \begin{matrix} n_p \\ -1 \end{matrix}$$

Once again, all the analysis of section 4 goes through with P referring to the matrix of eigenvectors corresponding to nonzero eigenvalues of H, $P^{\dagger}$ in place of P^{-1}, and the inverse of the smallest nonzero eigenvalue of KK^T in place of $\|(KK^T)^{-1}\|_2$.

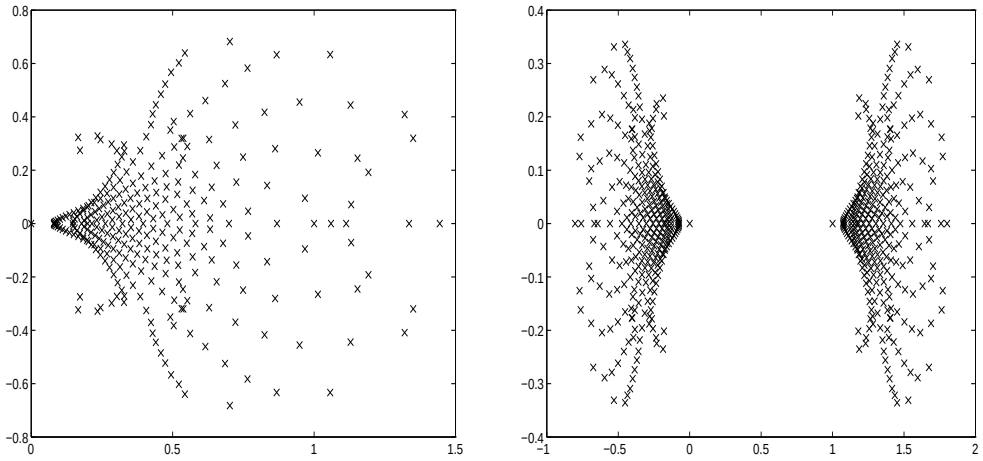


FIG. 1. *Eigenvalues of $\mathcal{A}_T$ (left) and $\mathcal{A}_D$ (right) for $\nu = 1/20$ and $n = 32$.*

6. Experimental results. We now present the results of numerical experiments that supplement the analysis of sections 3 through 5. Our test problem is a discrete Oseen operator (4) on $\Omega = (-1, 1) \times (-1, 1)$, with Dirichlet boundary conditions $u_1 = u_2 = 0$ on the three fixed walls ($x = -1$, $y = -1$, $x = 1$) and $u_1 = 1$, $u_2 = 0$ on the moving wall ($y = 1$). The coefficients of the convection terms describe a circular vortex

$$w_1 = 2y(1 - x^2), \qquad w_2 = -2x(1 - y^2).$$

We discretize using bilinear finite elements on a uniform rectangular $n \times n$ velocity grid of width $h = 2/n$ augmented by streamline upwinding [9, p. 185]. To impose div-stability, the pressure discretization uses "macro-elements" of width $2h$; see [8, p. 30]. The hydrostatic pressure is not explicitly specified, so that the matrices (3) are rank deficient by one. Additional details about this problem are given in [3]. All computations were performed in MATLAB on either a Sun SPARC-20 workstation or a DEC-Alpha 2100 4/275 workstation.

We first show in Figure 1 some sample distributions of eigenvalues of $\mathcal{A}_T$ and $\mathcal{A}_D$ for $\nu = 1/20$ and $n = 32$. The plot on the left gives an indication of the rectangle enclosing the eigenvalues of H (see (11)); the plot on the right represents the result of the mapping $\lambda \mapsto \left(1 \pm \sqrt{1 + 4\lambda}\right)/2$ of (37). Both pictures include the eigenvalue 1 of multiplicity $n_u = 2178$ for $\mathcal{A}_T$ and $n_u - n_p + 1 = 1890$ for $\mathcal{A}_D$.

We present our results primarily as tabulations of maximum perturbations and other quantities for various choices of viscosity parameter ν and grid parameter n. The rows and columns of the tables indicate behavior as either $\nu \to 0$ or n becomes large ($h \to 0$). Note that accurate discrete solutions to (4) are obtained only if ν is not too small relative to h. This difficulty can be ameliorated to some extent by an appropriate choice of discretization, such as the streamline upwinding method used here [2], [9, p. 262]. In practical experiments, it is often desired to compute solutions of a fixed accuracy for a variety of values of ν by letting $h \to 0$ and $\nu \to 0$ simultaneously. In an effort to follow trends in the data, we will consider some combinations of ν and

TABLE 1
Maximum normalized perturbations of eigenvalues of $\mathcal{A}_T$.

$n = 16$	ν	1	1/10	1/20	1/30	1/50	1/100
Original	$\lambda = 1$	21.7	137.6	318.5	478.2	766.1	1509.5
formulation	$\lambda \neq 1$	21.4	80.5	201.1	313.4	529.9	1110.0
Scaled	$\lambda = 1$	21.7	16.7	16.9	16.9	16.9	17.1
formulation	$\lambda \neq 1$	21.4	8.3	10.2	10.6	10.7	11.2
$n = 32$							
Original	$\lambda = 1$	15.7	58.6	56.6	83.0	172.0	402.4
formulation	$\lambda \neq 1$	22.7	26.7	50.8	43.1	141.2	257.8
Scaled	$\lambda = 1$	15.7	15.2	14.0	14.1	14.2	14.4
formulation	$\lambda \neq 1$	22.7	6.9	3.4	11.1	93.0	35.7

TABLE 2
Maximum normalized perturbations of eigenvalues of $\mathcal{A}_D$.

$n = 16$	ν	1	1/10	1/20	1/30	1/50	1/100
Original	$\lambda = 1$	13.4	13.5	13.7	14.2	14.8	15.8
formulation	$\mathrm{Re}(\lambda) < 0$	11.6	7.5	6.5	9.8	16.2	31.9
	$\mathrm{Re}(\lambda) > 1$	10.1	8.3	7.6	10.6	16.5	32.1
Scaled	$\lambda = 1$	13.4	13.5	13.7	14.2	14.8	15.8
formulation	$\mathrm{Re}(\lambda) < 0$	11.6	45.3	9.0	10.0	15.6	20.7
	$\mathrm{Re}(\lambda) > 1$	10.1	44.3	9.4	10.5	15.9	23.4
$n = 32$							
Original	$\lambda = 1$	12.9	12.9	13.0	13.0	13.1	13.7
formulation	$\mathrm{Re}(\lambda) < 0$	55.6	5.1	2.8	9.4	74.9	51.2
	$\mathrm{Re}(\lambda) > 1$	46.6	7.4	4.0	8.4	71.8	58.0
Scaled	$\lambda = 1$	12.9	12.9	13.0	13.0	13.1	13.7
formulation	$\mathrm{Re}(\lambda) < 0$	55.6	48.7	19.0	60.4	505.6	324.3
	$\mathrm{Re}(\lambda) > 1$	46.6	47.7	19.6	57.6	516.7	301.1

h that could produce inaccurate solutions.

Our main results are in Tables 1 and 2. Table 1 shows the effects of perturbation of the block tridiagonal preconditioner for two values of n and various ν. This data was obtained by computing the eigenvalues of a set of ten perturbed matrices $\mathcal{A}_T(\epsilon) = \mathcal{A}_T + \epsilon E$ where $\epsilon = 10^{-8}$ and E is a dense matrix with uniformly distributed random numbers in an interval $[-a_n, a_n]$, where $a_n = 16/n$. (For this choice, $\|E\|_2$ is approximately constant for all n.) For $\lambda = 1$, the table presents

$$\left(\max_{\hat{\lambda} \in \sigma_1(\mathcal{A}_T + \epsilon E)} |\hat{\lambda} - 1| \right) / \epsilon,$$

where $\sigma_1(\mathcal{A}_T + \epsilon E)$ is the set of n_p eigenvalues that are closest to 1, and the maximum is over all the perturbations. For $\lambda \neq 1$, the table shows

$$\left(\max_{\hat{\lambda} \in \sigma_1'(\mathcal{A}_T + \epsilon E)} |\hat{\lambda} - \lambda| \right) / \epsilon,$$

where $\sigma_1'(\mathcal{A}_T + \epsilon E)$ denotes the perturbations of $\lambda \in \sigma(\mathcal{A}_T)$, $\lambda \neq 1$, and $\lambda \neq 0$.[4]

[4] These were obtained by sorting $\sigma(\mathcal{A}_T)$ and $\sigma(\mathcal{A}_T + E)$ for each E and comparing the ordered sets. For all tests, similar results were obtained for other values of ϵ.

TABLE 3
Condition number of matrix P of eigenvectors of H.

ν	1	1/10	1/20	1/30	1/50	1/100
$n = 16$	5.4	18.0	5.4	5.0	7.9	14.7
$n = 32$	60.8	45.7	26.8	78.5	557.0	413.3
$n = 64$	181.8	3876.0	2900.4	2029.1	5441.2	2.74e4

TABLE 4
Norm of $(I - H)^{-1}$.

ν	1	1/10	1/20	1/30	1/50	1/100
$n = 16$	100.8	16.2	17.1	17.1	16.5	16.6
$n = 32$	70.8	70.9	23.4	16.7	43.2	17.3
$n = 64$	697.0	2.59e4	1124.9	318.7	61.9	231.1

TABLE 5
Secant of largest principal angle between Range(G^T) *and* Range(K^T) $(\alpha_{G,K})$.

ν	1	1/10	1/20	1/30	1/50	1/100
$n = 16$	1.11	1.65	2.51	3.29	4.53	6.51
$n = 32$	1.16	1.72	2.76	3.77	5.57	8.96
	xxx	xxx				

Analogous results for the block diagonal preconditioner are shown in Table 2. Here we also distinguish between the eigenvalues with positive and negative real parts.

We also present experimental results for three other quantities appearing in the bounds of sections 3 and 4: $\kappa(P)$, $\|(I - H)^{-1}\|_2$, and $\alpha_{G,K}$. These are shown in Tables 3, 4, and 5, respectively. Here $\kappa(P)$ refers to the version used for the rank-deficient problem as outlined in section 5. It was obtained by computing the matrix of eigenvectors P of H, normalizing the columns of P to have unit l_2-norm, and then computing the ratio of largest to smallest singular values of the submatrix of P corresponding to the nonzero eigenvalues of H. (The normalization ensures that we are within a factor of $\sqrt{n_p}$ of the condition number of the optimally scaled version of P [15].) Note that $\kappa(P)$ and $\alpha_{G,K}$ exhibit a general tendency to increase as either $\nu \to 0$ or n increases, with $\kappa(P)$ being particularly volatile.

Now consider the data of Table 1 for the block triangular preconditioner. Several trends are apparent.

1. In the original formulation, the perturbations of both $\lambda = 1$ and $\lambda \neq 1$ are increasing with $1/\nu$. In the scaled formulation, the perturbations of $\lambda = 1$ are insensitive to ν. This behavior is consistent with the results of Theorem 3.2.
2. In the scaled formulation, the perturbations of $\lambda \neq 1$ show some growth with $1/\nu$ for $n = 32$, although there is no clear trend. This may arise from growth in the product $\kappa(P)\,\|(I - H)^{-1}\|_2$.
3. There is little or no increase in perturbation size (and a decrease in some cases) with the change of grid size from $n = 16$ to $n = 32$. This contrasts with the bounds from the analysis, which degrade as $h \to 0$.

Next, consider Table 2.

1. The perturbations of $\lambda \neq 1$ for both the unscaled and scaled problems show some growth with $1/\nu$. The qualitative trends are similar, but the perturbations for the scaled formulations are larger. In contrast, the analysis (Theorem

TABLE 6
Maximum normalized perturbations of eigenvalues of H.

ν	1	1/10	1/20	1/30	1/50	1/100
$n = 16$	3.8	4.7	2.4	2.4	3.2	5.3
$n = 32$	8.3	4.8	3.4	10.8	73.5	31.1
$n = 64$	28.4	226.4	134.1	61.7	130.8	402.3

4.2) suggests that the scaled version would be smaller; therefore, it appears that the upper bounds of this analysis are not giving a complete indication of dependence on ν.

2. The perturbations of $\lambda = 1$ are insensitive to ν and there is essentially no difference between the perturbations of the eigenvalues with positive and negative real parts. As noted in Remark 4.1, we believe the dependence on $\alpha_{G,K}$ is an artifact of the proof.
3. The dependence on the mesh size is more pronounced for the triangular preconditioning, although there is no consistent pattern.

Thus, the analysis gives upper bounds on perturbation sizes, although it is not possible to completely correlate the analytic bounds and experimental results. This is likely due to the lack of a clear pattern in the behavior of $\kappa(P)$ and $\|(I - H)^{-1}\|_2$, together with the use of norm inequalities throughout the analysis that are not necessarily tight. For example, the analysis combines bounds on matrix l_2-norms derived from the underlying differential operator with l_∞ norms derived from Gerschgorin bounds. We suspect that the factor of $1/h$ used to relate these quantities artificially inflates the dependence of the bounds on the mesh size, although we see no way to avoid introducing this term. In general, both the analysis and experiments indicate a tendency for perturbations of the eigenvalues different from 1 to increase with $1/\nu$, but growth appears to be at worst linear and for the block triangular preconditioning it can be reduced by scaling.

Finally, Table 6 examines the sensitivity of the eigenvalues of the reduced matrix H to perturbation. The entries are

$$\left(\max_{\hat{\lambda} \in \sigma(H+\epsilon E)} |\hat{\lambda} - \lambda| \right) / \epsilon.$$

The results also indicate that the perturbations increase as the viscosity decreases, and they display some growth as the number of mesh points increase, roughly like that for the block diagonal preconditioner displayed in Table 2. Note that the dependence on both $1/\nu$ and n is much less severe than that of $\kappa(P)$ shown in Table 3. Once again, this stands in contrast to analytic bounds such as those obtainable from the Bauer–Fike theorem [7, p. 342], which suggests perturbations proportional to $\kappa(P)$.

7. Inexact inner iteration. We conclude by examining the effect of the inexact computation of the action of F^{-1}. For brevity, we restrict our attention to the unscaled version of $\mathcal{A}_T$; a similar analysis leads to essentially the same conclusions for $\mathcal{A}_D$. If the action of F^{-1} is approximated using an iterative method for each system $Fw = v$, then F^{-1} can be replaced by $F^{-1} + E$ in (9). The perturbed preconditioned matrix is then $\mathcal{A}_T + \mathcal{E}$, where

$$\mathcal{E} = \begin{pmatrix} FE & FEK^T \\ BE & BEK^T \end{pmatrix}.$$

For this analysis it will be useful to consider the complete version of (24),

$$\hat{\mathcal{E}}_T = \begin{pmatrix} FE(I + K^T(I-H)^{-1}G) & FEK^TP \\ P^{-1}(I-(I-H)^{-1})BE & P^{-1}(I-(I-H)^{-1})BEK^TP \\ \cdot(I + K^T(I-H)^{-1}G) & \end{pmatrix}.$$

Suppose the approximate solution $\tilde{w}$ satisfies

$$\frac{\|v - F\tilde{w}\|}{\|v\|} \leq \tau \tag{53}$$

for some tolerance τ. Standard inequalities yield the bounds on the relative error

$$\frac{\|w-\tilde{w}\|}{\|w\|} \leq \|F\|\,\|F^{-1}\|\,\tau, \qquad \frac{\|w-\tilde{w}\|}{\|w\|} \leq \|F\|\,\|E\|\,.$$

Treating these as approximate equalities gives $\|E\| \approx \tau\|F^{-1}\|$. It follows that

$$\begin{aligned}
\|\hat{E}_{11}\| &\lessapprox \tau\,\|F\|\,\|F^{-1}\|\,\left(1 + \|K^T\|\,\|(I-H)^{-1}\|\,\|G\|\right), \\
\|\hat{E}_{12}\| &\lessapprox \tau\,\|F\|\,\|F^{-1}\|\,\|K^T\|\,\|P\|\,, \\
\|\hat{E}_{21}\| &\lessapprox \tau\,\|P^{-1}\|\,(1 + \|(I-H)^{-1}\|)\,\|B\|\,\|F^{-1}\|\,\left(1 + \|K^T\|\,\|(I-H)^{-1}\|\,\|G\|\right), \\
\|\hat{E}_{22}\| &\lessapprox \tau\,\kappa(P)\,\|I - (I-H)^{-1}\|\,\|F^{-1}\|\,\|K^T\|.
\end{aligned}$$

Using the l_2-norm, let us consider the dependence of these bounds on the viscosity ν under the assumption (derived from the experimental results presented above) that the influence of P and H is not significant. The bounds on $\|\hat{E}_{11}\|$ and $\|\hat{E}_{21}\|$ include the product $\|K^T\|\,\|G\| = O(\nu)\cdot O(1/\nu)$, which is independent of ν. Using the coercivity and continuity of the convection–diffusion operator, it can be shown that as functions of ν,

$$\|F\| = O(1), \quad \|F^{-1}\| = O(1/\nu).$$

(An algebraic proof can be found in [4, Theorem 1].) This implies that the bounds on $\|\hat{E}_{11}\|$ and $\|\hat{E}_{21}\|$ grow as ν decreases, whereas the bounds on $\|\hat{E}_{12}\|$ and $\|\hat{E}_{22}\|$ are independent of ν. If τ is small, then as in section 3 we can restrict our attention to the block diagonal entries, which indicate that the perturbations of eigenvalues different from 1 in this particular case do not depend on ν. On the other hand, if τ is large, then it is not possible to exclude $\hat{E}_{21}$ from the Gerschgorin analysis, and the presence of $\|F^{-1}\|$ in this bound suggests that perturbations in all eigenvalues grow like $1/\nu$.

Consider the implication of these observations on the performance of iterative methods for solving the discrete Oseen equations (3). We demonstrated in [3] that the iteration counts of Krylov subspace methods such as GMRES with the preconditioner $\mathcal{Q}_T$ (7) are independent of the mesh size, but that there is some deterioration in performance as ν decreases. Now suppose an inner iteration with stopping criterion (53) is used to approximate the action of F^{-1}. The analysis given here suggests that if τ is large, there may be additional degradation of performance for small ν. In contrast, if τ is small then only $\lambda = 1$ is sensitive to perturbation, which suggests that (extra) degradation of the inner iteration with decreasing ν may not be as pronounced. Figure 2 shows the results of numerical experiments that corroborate these observations. The figure plots iteration counts of right-preconditioned "flexible" GMRES (FGMRES)

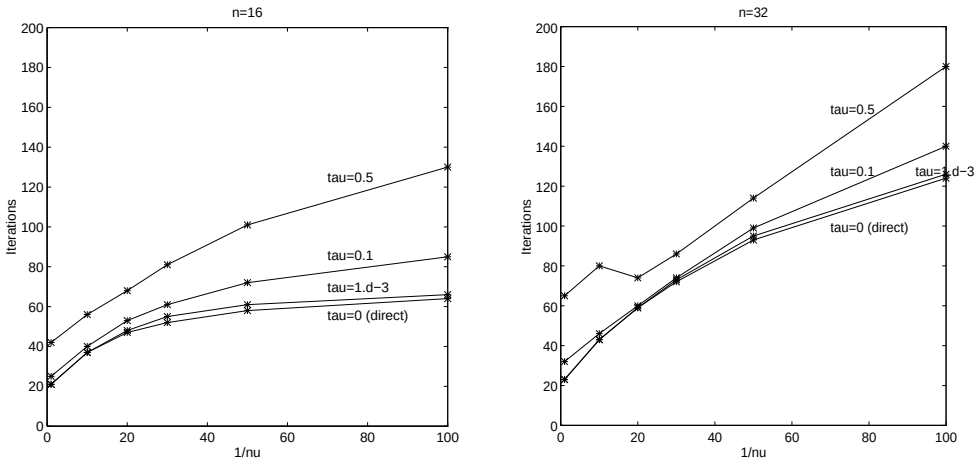


FIG. 2. *Iterations of FGMRES with inexact convection–diffusion solves.*

[12] as a function of ν for solving the discrete problem (3). The inner iteration for the convection–diffusion subproblems $Fw = v$ was a line Gauss–Seidel method with stopping criterion (53).[5] The test matrices were those used in section 6; the right-hand side f consisted of normally distributed random numbers with mean 0 and variance 1. The outer iteration used a zero initial guess and was stopped when the relative residual in the Euclidean norm was less than or equal to 10^{-6}. The results with inner iteration are compared with the use of a direct method for the action of F^{-1}. They indicate that for the relatively modest tolerance $\tau = 10^{-3}$, the inexact inner solves lead to little increase in outer iterations for any ν. For less stringent τ, additional outer iterations are required and the number of additional iterations becomes larger as ν decreases.

Remark 7.1. It can be shown that for small τ the perturbations of the eigenvalues of the scaled system behave in the same way as those for the unscaled system. However, scaling affects the relative weighting given to the two block equations of (3), which in turn may affect the iterative solver. Therefore, we have restricted our attention here to the unscaled system.

Acknowledgments. The author acknowledges helpful discussions with Santiago Arteaga, Luca Pavarino, and Pete Stewart and insightful comments from Hans Weinberger. The numerical results with FGMRES used software produced by Tim Kelley [11].

REFERENCES

[1] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[2] A. BROOKS AND T. HUGHES, *Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations*, Comp. Meth. Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[5] FGMRES allows the preconditioner to vary during the course of the iteration; it was used here because the number of inner iterations was not generally constant; i.e., the preconditioner is not a fixed linear operator. FGMRES is equivalent to GMRES when a direct solve is used for F^{-1}.

[3] H. Elman and D. Silvester, *Fast nonsymmetric iterations and preconditioning for Navier–Stokes equations*, SIAM J. Sci. Comput., 17 (1996), pp. 33–46.
[4] H. C. Elman and M. H. Schultz, *Preconditioning by fast direct methods for nonselfadjoint nonseparable elliptic problems*, SIAM J. Numer. Anal., 23 (1986), pp. 44–57.
[5] R. Freund and N. M. Nachtigal, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.
[6] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, New York, 1984.
[7] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[8] M. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows*, Academic Press, San Diego, CA, 1989.
[9] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, New York, 1987.
[10] O. A. Karakashian, *On a Galerkin–Lagrange multiplier method for the stationary Navier–Stokes equations*, SIAM J. Numer. Anal., 19 (1982), pp. 909–923.
[11] T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*. SIAM, Philadelphia, PA, 1995.
[12] Y. Saad, *A flexible inner–outer preconditioned GMRES algorithm*, SIAM J. Sci. Comput., 14 (1993), pp. 461–469.
[13] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.
[14] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, 2nd ed., Springer-Verlag, New York, 1993.
[15] A. van der Sluis, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969), pp. 14–23.
[16] R. Verfürth, *A combined conjugate gradient-multigrid algorithm for the numerical solution of the Stokes problem*, IMA J. Numer. Anal., 4 (1984), pp. 441–455.
[17] A. J. Wathen, *Realistic eigenvalue bounds for the Galerkin mass matrix*, IMA J. Numer. Anal., 7 (1987), pp. 449–457.
[18] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, UK, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 3, pp. 752–774, July 1997

012

EXTENSION OF ISOMETRIES IN FINITE-DIMENSIONAL INDEFINITE SCALAR PRODUCT SPACES AND POLAR DECOMPOSITIONS*

YURI BOLSHAKOV†, CORNELIS V. M. VAN DER MEE‡, ANDRÉ C. M. RAN§, BORIS REICHSTEIN¶, AND LEIBA RODMAN‖

Abstract. Witt's theorem on the extension of H-isometries to H-unitary matrices with respect to the scalar product generated by a self-adjoint nonsingular matrix H is studied in detail. All possible extensions are given, and their structure as a real analytic manifold is described. Analogous problems with respect to skew-symmetric scalar products are studied as well.

The main motivation to study these problems, as well as the main applications of the results obtained, concerns polar decompositions in indefinite scalar product spaces. As another application, for given B all solutions of the matrix equation $XA = B$ with H-unitary X and upper triangular A are described. Equations of this type are of vital importance in hyperbolic QR decompositions.

Key words. indefinite scalar products, isometries, polar decompositions, hyperbolic QR decompositions

AMS subject classifications. 15A63, 15A23

PII. S0895479895290644

1. Introduction.

Let F be either the field of real numbers $\mathbf{R}$ or the field of complex numbers $\mathbf{C}$. Fix a real symmetric (if $F = \mathbf{R}$) or complex Hermitian (if $F = \mathbf{C}$) invertible $n \times n$ matrix H. Consider the scalar product induced by H according to the formula $[x, y] = \langle Hx, y\rangle$, $x, y \in F^n$. Here $\langle \cdot\,, \cdot \rangle$ stands for the standard scalar product in F^n defined by $\langle x, y\rangle = \sum_{j=1}^{n} x_j \bar{y}_j$, where $(x_1, \ldots, x_n)^T$ and $(y_1, \ldots, y_n)^T$ are column vectors in F^n. (Of course, $\bar{y}_j = y_j$ if $F = \mathbf{R}$.) The scalar product $[\cdot\,, \cdot]$ is nondegenerate ($[x, y] = 0$ for all $y \in F^n$ implies $x = 0$) but is indefinite in general. In other words, the real number $[x, x]$ can be positive, negative, or zero for various $x \in F^n$ (unless H is definite). The vector $x \in F^n$ is called *positive* if $[x, x] > 0$, *neutral* if $[x, x] = 0$, and *negative* if $[x, x] < 0$.

Well-known concepts related to the scalar product $[\cdot\,, \cdot]$ are defined in obvious ways. Thus, given an $n \times n$ matrix A over F, the adjoint $A^{[*]}$ is defined by $[Ax, y] = [x, A^{[*]}y]$ for all $x, y \in F^n$. The formula $A^{[*]} = H^{-1}A^*H$ is verified immediately. (Here and elsewhere we denote by A^* the conjugate transpose of A, so that $A^* = A^T$ if $F = \mathbf{R}$.) A matrix A is called *H-self-adjoint* if $A^{[*]} = A$ or, equivalently, if HA is Hermitian. An $n \times n$ matrix U is called *H-unitary* if $[Ux, Uy] = [x, y]$ for all $x, y \in F^n$

*Received by the editors August 21, 1995; accepted for publication (in revised form) by P. Lancaster August 20, 1996.

http://www.siam.org/journals/simax/18-3/29064.html

†Department of Mathematics, Yaroslavl State University, Yaroslavl, Russia.

‡Dipartimento di Matematica, Università di Cagliari, Via Ospedale 72, 09124 Cagliari, Italy. The work of this author was performed under the auspices of C. N. R.-G. N. F. M. and was partially supported by the research project "Nonlinear problems in analysis and its physical, chemical and biological applications: Analytical, modelling and computational aspects" of the Italian Ministry of Higher Education and Research (M. U. R. S. T.).

§Faculteit Wiskunde en Informatica, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands (ran@cs.vu.nl).

¶Department of Mathematics, The Catholic University of America, Washington, DC 20064.

‖Department of Mathematics, The College of William and Mary, Williamsburg, VA 23187-8795 (lxrodm@mail.wm.edu). The work of this author was partially supported by NSF grant DMS 9500924.

or, equivalently, $U^*HU = H$. Observe that for every H-unitary matrix U we have $|\det U| = 1$; in particular, $\det U = \pm 1$ if $F = \mathbf{R}$.

This article is the third of a series of four articles on decompositions of an $n \times n$ matrix X over F of the form

$$X = UA, \tag{1.1}$$

where U is H-unitary and A is H-self-adjoint (with or without additional restrictions). We call the decomposition (1.1) an *H-polar decomposition* of X. Our first article, henceforth called [BMRRR1], is devoted to the existence, uniqueness (up to equivalence), and basic properties of decompositions (1.1) and to the existence of H-polar decompositions of H-normal matrices. In our second article, subsequently referred to as [BMRRR2], we studied decompositions of the type (1.1), where various constraints are imposed on the matrices X, U, A, and H, and discussed their applications in linear optics.

In studying H-polar decompositions, we often face the problem of extending H-isometries between linear subspaces to H-isometries defined on the whole space. The theorem stating the existence of such extensions is a classical result in geometry called Witt's theorem (see, e.g., [A, Theorem III.3.9.], or [D]). However, the classical results are concerned with the existence of a Witt extension and do not address the problems of listing all possible Witt extensions and describing their topological and algebraic structure. In the present paper, we give a detailed proof of Witt's theorem in both the real and the complex cases, detailed enough to yield all Witt extensions that exist. This is the content of section 2. As a by-product, the connected components of the set of all Witt extensions are described in section 3. Section 4 is devoted to the analogous problem of finding real Witt extensions with respect to a real skew-symmetric scalar product.

Another aspect of the present paper concerns a particular class of H-polar decompositions (1.1) in which the matrix A is H-nonnegative (i.e., HA is positive definite Hermitian). Such decompositions will be called *semidefinite H-polar decompositions*. In section 5 the semidefinite H-polar decompositions are described and characterized in full detail using the general results of [BMRRR1] as a starting point and applying the results on Witt extensions of section 2.

In section 6 the description of all Witt extensions is applied to a class of matrix decompositions, namely, hyperbolic QR decompositions, which are crucial in certain algorithms based on the generalized Schur method (see, e.g., [B, OSB, V]).

We remark in passing that the results of this paper concerning the description of Witt's extensions carry over to certain fields other than $\mathbf{R}$ or $\mathbf{C}$. Indeed, our description involves H-unitary matrices; normalization of vectors needed to construct such matrices is only possible in number fields closed with respect to the square root operation on positive numbers, such as the field of real algebraic numbers.

The following notation will be used. The number of positive (negative, zero) eigenvalues of a Hermitian matrix A is denoted by $\pi(A)$ ($\nu(A), \delta(A)$). The symbol F^n (where $F = \mathbf{R}$ or $F = \mathbf{C}$) stands for the vector space of n-dimensional columns over F. We denote by $F^{m\times n}$ the vector space of $m \times n$ matrices over F. I_m is the $m \times m$ identity matrix. The block diagonal matrix with matrices $Z_1, \dots, Z_k$ on the main diagonal is denoted by $Z_1 \oplus \cdots \oplus Z_k$ or $\operatorname{diag}(Z_1, \dots, Z_k)$. The set of eigenvalues (including nonreal eigenvalues for real matrices) of a matrix X is denoted by $\sigma(X)$. A^T stands for the transpose of a matrix A. $\operatorname{Ker} A$ and $\operatorname{Im} A$ stand for the null space and range of a matrix A. The symbol $\mathcal{M} \oplus \mathcal{N}$ denotes the direct sum of the subspaces

$\mathcal{M}$ and $\mathcal{N}$. For a subspace $\mathcal{M} \subset F^n$ and an indefinite scalar product $[x, y] = \langle Hx, y\rangle$, we call the subspace

$$\mathcal{M}^{[\perp]} = \{x \in F^n | [x, y] = 0 \text{ for all } y \in \mathcal{M}\}$$

the *H-orthogonal companion* of $\mathcal{M}$.

2. Witt's theorem and its refinements. In this section we will derive a version of Witt's theorem which is suitable to our framework and describe all H-isometries to which a given partial H-isometry can be extended.

We start with Witt's theorem, which is a classical result (see, e.g., [A, D]). The proofs given in [A, D] are algebraic and do not easily yield the parametrization that we need. Although the proofs from [A, D] could be adapted, doing so would create a portion of the paper at odds in style with the linear algebra methods of the rest of the paper. On the other hand, results on extensions of isometries are well known in the theory of operators in infinite-dimensional spaces with indefinite scalar products; see, e.g., section 5.2 in [AI1] or section II.9 in [IKL].

THEOREM 2.1. *Let* $[\,\cdot\,,\,\cdot\,]_1$ *and* $[\,\cdot\,,\,\cdot\,]_2$ *be the two scalar products in* F^n *defined by the invertible Hermitian* $n \times n$ *matrices* H_1 *and* H_2, *respectively*:

$$[x, y]_1 = \langle H_1 x, y\rangle, \quad [x, y]_2 = \langle H_2 x, y\rangle, \qquad x, y \in F^n.$$

Assume $\pi(H_1) = \pi(H_2)$. *Let* $U_0 : V_1 \to V_2$, *where* V_1 *and* V_2 *are subspaces in* F^n, *be a nonsingular linear transformation that preserves the scalar products*

$$[U_0 x, U_0 y]_2 = [x, y]_1 \quad \textit{for every} \quad x, y \in V_1. \tag{2.1}$$

Then there exists a linear transformation $U : F^n \to F^n$ *such that*

$$[Ux, Uy]_2 = [x, y]_1 \quad \textit{for every} \quad x, y \in F^n \tag{2.2}$$

and

$$Ux = U_0 x \quad \textit{for every} \quad x \in V_1. \tag{2.3}$$

It is easy to see that the condition $\pi(H_1) = \pi(H_2)$, the nonsingularity of U_0, and the equality (2.1) are necessary for the existence of U with the asserted properties. Note that any such U is necessarily invertible; however, a linear transformation U_0 that satisfies (2.1) need not be invertible. A linear transformation (or its matrix representation with respect to specified bases) U with the property (2.2) is called *H_1-H_2-unitary*.

Given U_0 as in Theorem 2.1, any linear transformation U satisfying (2.2) and (2.3) will be called a *Witt extension* of U_0.

The following well-known fact will be useful in the proof of Theorem 2.1.

PROPOSITION 2.2. *Let* $[x, y] = \langle Hx, y\rangle$ *be an indefinite scalar product on* F^n. *The following statements are equivalent for the subspace* $\mathcal{M} \subset F^n$:

(i) $\mathcal{M}$ *is H-nondegenerate; i.e.,* $x_0 \in \mathcal{M}$, $[x_0, y] = 0$ *for all* $y \in \mathcal{M}$ *implies* $x_0 = 0$.

(ii) *The H-orthogonal companion* $\mathcal{M}^{[\perp]}$ *is H-nondegenerate.*

(iii) $\mathcal{M}^{[\perp]}$ *is a direct complement to* $\mathcal{M}$ *in* F^n.

The proof is based on the simple observation that $\dim \mathcal{M} + \dim \mathcal{M}^{[\perp]} = n$; see [GLR] or [Bo] for complete details.

Proof of Theorem 2.1. We give a proof of Theorem 2.1 which will also serve as a basis for subsequent results concerning detailed descriptions of all Witt extensions. Put $m = m_+ + m_- + m_0$. Let $\{e_i\}_{i=1,2,\dots,m}$ be a basis of V_1 such that $[e_j, e_j]_1 = 1$ for $j = m_0 + 1, m_0 + 2, \dots, m_0 + m_+$, $[e_k, e_k]_1 = -1$ for $k = m_0 + m_+ + 1,\ m_0 + m_+ + 2, \dots, m$, and all the remaining indefinite scalar products of the basis vectors are zero (thus the Hermitian matrix defining the H_1-scalar product on V_1 has m_+ positive eigenvalues and m_- negative eigenvalues and the multiplicity of zero is m_0). Introduce the m linear functionals α_i on F^n as follows:

$$\alpha_i(x) = [x, e_i]_1, \qquad i = 1, 2, \dots, m.$$

Since $\alpha_1, \dots, \alpha_m$ are linearly independent, there exist vectors $\tilde{e}_i \in F^n$ such that $\alpha_i(\tilde{e}_j) = \delta_{ij}$, where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$, i.e., such that $[e_i, \tilde{e}_j]_1 = \delta_{ij}$ for all $i, j = 1, 2, \dots, m$. Let

$$W_k = \operatorname{span}\{e_k, \tilde{e}_k\}, \quad k = 1, 2, \dots, m_0.$$

Since $[e_k, e_k]_1 = 0$ and $[e_k, \tilde{e}_k]_1 = 1$, each W_k is H_1-nondegenerate. Without loss of generality we can assume that, for $k = 1, 2, \dots, m_0$, we have $[\tilde{e}_k, \tilde{e}_k]_1 = 0$. (Indeed, we can always replace the vector $\tilde{e}_k$ by the vector $\tilde{e}_k - \frac{1}{2}[\tilde{e}_k, \tilde{e}_k]_1 e_k$, which has the above property.) Let

$$e'_k = \frac{1}{\sqrt{2}}(e_k - \tilde{e}_k), \qquad e''_k = \frac{1}{\sqrt{2}}(e_k + \tilde{e}_k).$$

It is easy to see that

$$[e'_k, e'_k]_1 = -1, \quad [e''_k, e''_k]_1 = 1, \quad [e'_k, e''_k]_1 = 0.$$

The subspace $W = W_1 + \dots + W_{m_0} + \operatorname{span}\{e_j\}_{i=m_0+1,\dots,m}$ is H_1-nondegenerate; hence, $W^{[\perp]}$ is H_1-nondegenerate by Proposition 2.2. Therefore, we can append the vectors

$$e_s, \quad s = 2m_0 + m_+ + m_- + 1,\ 2m_0 + m_+ + m_- + 2, \dots, n$$

to the set

$$\{e'_k, e''_k, e_{m_0+1}, e_{m_0+2}, \dots, e_m\}_{k=1,2,\dots,m_0}$$

of $2m_0 + m_+ + m_-$ vectors such that the resulting ordered set $\{g_1, \dots, g_n\}$ will be a basis in F^n with the property that $[g_i, g_j]_1 = \epsilon_i \delta_{ij}$ for $i, j = 1, \dots, n$, where $\epsilon_i = \pm 1$.

Now let $f_i = U_0 e_i$, $i = 1, 2, \dots, m$. We introduce vectors f'_k and f''_k ($k = 1, 2, \dots, m$) and vectors f_s ($s = 2m_0 + m_+ + m_- + 1,\ 2m_0 + m_+ + m_- + 2, \dots, n$) in the same way we introduced the vectors e'_k, e''_k, and e_s but using $[\,\cdot\,,\,\cdot\,]_2$ instead of $[\,\cdot\,,\,\cdot\,]_1$, resulting in a basis $h_1, \dots, h_n$ in F^n. The hypotheses on H_1 and H_2 ($\pi(H_1) = \pi(H_2)$) and on U_0 (U_0 being an isometry) guarantee that $[h_i, h_j]_2 = [g_i, g_j]_1$ $(i, j = 1, \dots, n)$.

Define the $n \times n$ matrix U by the equalities

$$Ue'_k = f'_k,\ Ue''_k = f''_k, \quad k = 1, 2, \dots, m_0,$$

$$Ue_s = f_s, \qquad s = 2m_0 + 1,\ 2m_0 + 2, \dots, n.$$

It is easy to see that the matrix U has all the properties required by the statement of the theorem. □

We will use the bases

$$\mathcal{E} = \{e_1, e_2, \dots, e_{m_0}, e_{m_0+1}, \dots, e_m, \tilde{e}_1, \dots, \tilde{e}_{m_0}, e_{2m_0+m_++m_-+1}, \dots, e_n\} \tag{2.4}$$

and $\mathcal{F}$ (consisting of the vectors Ue, where $e \in \mathcal{E}$) of F^n constructed in the proof of Theorem 2.1. These bases will be more convenient than the ones we considered above because the subspaces V_1 and V_2 are spanned by the first m vectors of the corresponding bases. Recall that $U_0V_1 = V_2$, as U_0 is nonsingular. Thus, in particular, $\dim V_1 = \dim V_2$. With respect to $[\,\cdot\,,\,\cdot\,]_1$, the basis (2.4) has the Gramian matrix

$$\begin{bmatrix} 0 & 0 & I & 0 \\ 0 & J_1 & 0 & 0 \\ I & 0 & 0 & 0 \\ 0 & 0 & 0 & J_2 \end{bmatrix}, \tag{2.5}$$

where I is the $m_0 \times m_0$ identity matrix and J_1 is the diagonal $(m_+ + m_-) \times (m_+ + m_-)$ matrix such that its first m_+ diagonal elements are $+1$ and its remaining m_- diagonal elements are -1. Similarly, J_2 is the Gramian matrix of the basis $\{e_{m_0+m+1}, e_{m_0+m+2}, \dots, e_n\}$ of the subspace spanned by these vectors; without loss of generality we can (and do) assume that J_2 is a diagonal matrix for which several diagonal entries are $+1$ and the remaining diagonal entries are -1.

The matrix (2.5) is also the Gramian matrix of the basis $\mathcal{F}$ with respect to $[\,\cdot\,,\,\cdot\,]_2$. The matrix U (constructed in the proof of Theorem 2.2), when understood as a linear transformation $F^n \to F^n$, is the $n \times n$ identity matrix with respect to the basis $\mathcal{E}$ (in F^n as the domain space of U) and the basis $\mathcal{F}$ (in F^n as the image space of U).

The Witt extensions of a given U_0 are described by the following theorem. (We represent the Witt extensions as linear transformations $F^n \to F^n$ with respect to the bases $\mathcal{E}$ and $\mathcal{F}$ constructed above.)

THEOREM 2.3 (extended Witt's theorem). *If a matrix $\tilde{U}$ is a Witt extension of the matrix U_0, then there exist a J_2-unitary matrix P_1 (of order $n - m - m_0$), an $(n - m - m_0) \times m_0$ matrix P_2, and a skew-self-adjoint $m_0 \times m_0$* matrix P_3 (*i.e., $P_3^* = -P_3$) such that the matrix of $\tilde{U}$ has the form*

$$\tilde{U} = \begin{bmatrix} I_{m_0} & 0 & -\frac{1}{2}P_2^*J_2P_2 + P_3 & -P_2^*J_2P_1 \\ 0 & I_{m-m_0} & 0 & 0 \\ 0 & 0 & I_{m_0} & 0 \\ 0 & 0 & P_2 & P_1 \end{bmatrix}. \tag{2.6}$$

Here $m = \dim V_1$ and m_0 is the number of zero eigenvalues of the Gramian matrix of any basis in V_1 with respect to $[\,\cdot\,,\,\cdot\,]_1$.

Conversely, if P_1 is an arbitrary J_2-unitary matrix, P_2 is an arbitrary $(n - m - m_0) \times m_0$ matrix, and P_3 is an arbitrary skew-self-adjoint $m_0 \times m_0$ matrix, then the matrix $\tilde{U}$ defined by (2.6) *is a Witt extension of U_0.*

Proof. The proof is straightforward. Any extension $\tilde{U}$ of U_0 in the above bases has the matrix

$$\tilde{U} = \begin{bmatrix} I & 0 & A_1 & A_2 \\ 0 & I & A_3 & A_4 \\ 0 & 0 & A_5 & A_6 \\ 0 & 0 & A_7 & A_8 \end{bmatrix}. \tag{2.7}$$

The necessary and sufficient condition for the matrix (2.7) to be H_1-H_2-unitary is the identity $H_1^{-1}\tilde{U}^*H_2\tilde{U} = I$. Taking into account (2.5) and (2.7) we can rewrite the last

relation in block form as

$$\begin{bmatrix} A_5^* & A_3^*J_1 & u_{13} & u_{14} \\ 0 & I & A_3 & A_4 \\ 0 & 0 & A_5 & A_6 \\ J_2A_6^* & J_2A_4^*J_1 & u_{43} & u_{44} \end{bmatrix} = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix}, \tag{2.8}$$

where

$$\begin{aligned} u_{13} &= A_5^*A_1 + A_3^*J_1A_3 + A_1^*A_5 + A_7^*J_2A_7, \\ u_{14} &= A_5^*A_2 + A_3^*J_1A_4 + A_1^*A_6 + A_7^*J_2A_8, \\ u_{43} &= J_2A_6^*A_1 + J_2A_4^*J_1A_3 + J_2A_2^*A_5 + J_2A_8^*J_2A_7, \\ u_{44} &= J_2A_6^*A_2 + J_2A_4^*J_1A_4 + J_2A_2^*A_6 + J_2A_8^*J_2A_8. \end{aligned}$$

Equating the corresponding blocks in (2.8), we derive the theorem statement. □

If V_1 is H_1-nondegenerate (i.e., $m_0 = 0$), then necessarily V_2 is H_2-nondegenerate and the result of Theorem 2.3 is obvious.

Observe also that the inverse of the matrix (2.6) is given by

$$\tilde{U}^{-1} = \begin{bmatrix} I & 0 & -\frac{1}{2}\hat{P}_2^*J_2\hat{P}_2 + \hat{P}_3 & -\hat{P}_2^*J_2\hat{P}_1 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & \hat{P}_2 & \hat{P}_1 \end{bmatrix}, \tag{2.9}$$

where

$$\hat{P}_1 = P_1^{-1} = J_2P_1^*J_2,\ \hat{P}_2 = -P_1^{-1}P_2 = -J_2P_1^*J_2P_2,\ \hat{P}_3 = -P_3.$$

Note that a Witt extension $\tilde{U}$ has the form (2.6) with respect to different bases in domain and image space, namely, with respect to the basis $\mathcal{E}$ given by (2.4) in the domain space and the basis $\mathcal{F}$ consisting of the vectors Ue in the image space. Keeping this in mind, we can easily reformulate Theorem 2.3 in the following way with respect to one basis $\mathcal{E}$ (the same for both the domain and the range of $\tilde{U}$). It is this form of the theorem that we shall apply later in Theorems 5.6 and 6.1.

THEOREM 2.4 (extended Witt's theorem, second version). *Let U be a fixed Witt extension of U_0 as constructed in Theorem* 2.1. *Then any Witt extension $\tilde{U}$ of U_0 is given by $\tilde{U} = UM$, where M has the form of the right-hand side of* (2.6) *with respect to the basis $\mathcal{E}$ in* (2.4).

Proof. Observe that U maps the elements of the basis $\mathcal{E}$ into the corresponding elements of the basis $\mathcal{F}$ and that (2.6) is the matrix representation of U with respect to the basis $\mathcal{E}$ in the domain space and the basis $\mathcal{F}$ in the image space. □

It is of interest to compute the number of independent real parameters that describe all Witt extensions. Assume first $F = \mathbf{C}$. Then the formula (2.6), combined with the real analytic description of the group of J_2-unitary matrices (see, e.g., Theorem IV.3.1 in [GLR]), produces the following result.

THEOREM 2.5 ($F = \mathbf{C}$). *The set $W(U_0)$ of all Witt extensions of a given isometry $U_0 : V_1 \to V_2$ is parametrized by $(n-m)^2$ independent real variables, where $m =$* dim V_1. *More precisely, let*

$$p = \pi(H_1) - m_+ - m_0, \qquad q = \nu(H_1) - m_- - m_0, \tag{2.10}$$

where m_+, m_-, and m_0 are the numbers of positive, negative, and zero eigenvalues, respectively, of the Gramian matrix of any basis in V_1 with respect to $[\,\cdot\,,\,\cdot\,]_1$. Then $W(U_0)$ is diffeomorphic (as a real analytic manifold) to

$$SU(p)\times SU(q)\times T\times T\times \mathbf{R}^w, \qquad w=2pq+2(n-m-m_0)m_0+m_0^2$$

if both p and q are positive, and $W(U_0)$ is diffeomorphic to

$$SU(p+q)\times T\times \mathbf{R}^w$$

if exactly one of p and q is zero. Finally, $W(U_0)$ is diffeomorphic to $\mathbf{R}^w$ if $p=q=0$. Here T is the unit circle and

$$SU(k)=\{X\in \mathbf{C}^{k\times k}|\ X \text{ unitary, } \det X=1\}$$

is the $k\times k$ special unitary group.

Proof. We use the notation of Theorem 2.3. The matrix $\tilde{U}$ is parametrized by (P_1,P_2,P_3), where P_2 and P_3 are in turn parametrized by $2(n-m-m_0)m_0$ and m_0^2 independent real variables, respectively. Observe that $p=\pi(J_2)$ and $q=\nu(J_2)$. Thus, the group of all J_2-unitary matrices is diffeomorphic (as a real analytic manifold) either to $SU(p)\times SU(q)\times T\times T\times \mathbf{R}^{2pq}$ (if both p and q are positive) or to $SU(p+q)\times T$ (if exactly one of p and q is zero); see, e.g., Theorem IV.3.1 in [GLR]. In fact, explicit charts for the group of all J_2-unitary matrices can be constructed using the diffeomorphism mentioned above and the following two charts for $SU(p)$, namely, the sets

$$\left\{\exp K\,:\,K=-K^*,\ \text{trace } K=0,\ \sigma(K)\subset \pm\left(-\frac{\pi}{2}i,\frac{3\pi}{2}i\right]\right\}.$$

The number of real parameters describing the group of J_2-unitary matrices is $(p^2-1)+(q^2-1)+1+1+2pq=(p+q)^2$ if $p,q>0$. (Here we use the fact that $SU(k)$ has real dimension k^2-1, equal to the real dimension of the set of all skew-self-adjoint $k\times k$ matrices with trace 0, which is the Lie algebra of $SU(k)$.) The group of J_2-unitary matrices has real dimension $(p+q)^2$ also in the case where exactly one of p and q is zero. Thus, the total number of real parameters describing $\tilde{U}$ is

$$\begin{aligned}(p+q)^2+2(n-m-m_0)m_0+m_0^2&=(p+q+m_0)^2\\&=(\pi(H_1)+\nu(H_1)-m_+-m_--m_0)^2\\&=(n-m)^2.\end{aligned}$$

An analogous proof also works in the case $p=q=0$. ∎

The real analogue of Theorem 2.5 runs as follows.

THEOREM 2.6 ($F=\mathbf{R}$). *Let $m=\dim V_1$, and let p and q be defined by (2.10). Then the set $W(U_0)$ of all Witt extensions of an isometry $U_0: V_1\to V_2$ is connected if $p=q=0$, has two connected components if exactly one of p and q is positive, and has four connected components if both p and q are positive. Every connected component of $W(U_0)$ is diffeomorphic (as a real analytic manifold) to*

$$SO(p)\times SO(q)\times \mathbf{R}^v, \qquad v=pq+(n-m-m_0)m_0+\frac{1}{2}m_0(m_0-1),$$

where $SO(k)$ is the group of real unitary (i.e., real orthogonal) $k\times k$ matrices with determinant 1 if both p and q are positive; every connected component of $W(U_0)$ is diffeomorphic to

$$SO(p+q)\times \mathbf{R}^v$$

if exactly one of p and q is zero. Finally, $W(U_0)$ is diffeomorphic to $\mathbf{R}^v$ if $p = q = 0$. In all cases, every connected component of $W(U_0)$ can be parametrized by $\frac{1}{2}(n-m)(n-m-1)$ independent real variables.

The part of Theorem 2.6 concerning the number of connected components follows immediately from Theorems 2.3 and 3.1. (The latter is stated and proved in the next section.) The remainder of the proof of Theorem 2.6 is analogous to that of Theorem 2.5: one should use the real analogue of Theorem IV.3.1 in [GLR] and the fact that $SO(k)$ has (real) dimension $\frac{1}{2}k(k-1)$; this is the dimension of the Lie algebra of $SO(k)$ which consists of all real skew-symmetric $k \times k$ matrices.

It is a curious observation that the number of real parameters describing $W(U_0)$ depends only on n (the order of H_1) and on m (the dimension of V_1) and does not depend on m_0 (the degree of degeneracy of V_1 in the indefinite scalar product induced by H_1).

In particular, Theorems 2.5 and 2.6 allow one to identify the fundamental group of the set $W(U_0)$ using the well-known fact that $SU(k)$ and $\mathbf{R}^n$ are simply connected; the fundamental group of $SO(k)$ is of order 2 if $k \geq 3$, the infinite cyclic group $\mathbf{Z}$ if $k = 2$, and the trivial group if $k = 1$; and the fundamental group of the product of two arcwise connected topological spaces X and Y is the direct product of the fundamental groups of X and Y (see, e.g., sections II.VIII, II.X, and II.XI in [C]). Thus the fundamental group of $W(U_0)$ is $G_p \times G_q$ if both p and q are positive, G_{p+q} if one of p, q is positive and the other vanishes, and trivial if $p = q = 0$; here $G_p = \mathbf{Z}$ if $F = \mathbf{C}$, whereas $G_p = \mathbf{Z}_2$ if $p \geq 3$, $G_2 = \mathbf{Z}$, and G_1 is trivial if $F = \mathbf{R}$.

We conclude this section with two illustrative examples.

Example 2.1. Let $H = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$; $V = \operatorname{span}\left\{\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right\}$. Any linear transformation $U_0 : V \to V$ is an isometry. The linear transformation $U_0 : V \to V$ is defined by $U_0\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}$, where $\alpha \neq 0$ is a given complex number. We shall find the Witt extensions U of U_0. An elementary calculation shows that all such U have the form $\begin{bmatrix} \alpha & x \\ 0 & \bar{\alpha}^{-1} \end{bmatrix}$, where $x \in \mathbf{C}$ is any number such that $\bar{\alpha}x + \bar{x}\alpha = 0$. If we consider $F = \mathbf{R}$, then α is real and the unique Witt extension of U_0 is given by $\operatorname{diag}(\alpha, \alpha^{-1})$. □

Example 2.2. Let $H = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$; $V = \operatorname{span}\left\{\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right\}$. A linear transformation $U_0 : V \to V$ defined by $U_0\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}$, $\alpha \neq 0$, is an H-isometry if and only if $|\alpha| = 1$. The Witt extensions U of U_0 are described by $U = \operatorname{diag}(\alpha, y)$, where $|y| = 1$. In the real case we have exactly *two* Witt extensions (corresponding to $y = \pm 1$). □

3. Connectivity of the H-unitary groups. Let H be an invertible Hermitian $n \times n$ matrix over F ($F = \mathbf{R}$ or $F = \mathbf{C}$). The set of H-unitary matrices (over F) is easily seen to be a group, denoted $\mathcal{U}(H; F)$. Its connected components are described as follows.

THEOREM 3.1.

(a) *The group $\mathcal{U}(H; \mathbf{C})$ is connected.*

(b) *If $F = \mathbf{R}$ and H is definite (positive or negative), then the group $\mathcal{U}(H; \mathbf{R})$ has two connected components. One of them contains all $X \in \mathcal{U}(H; \mathbf{R})$ with* $\det X = 1$; *the other contains all $X \in \mathcal{U}(H; \mathbf{R})$ with* $\det X = -1$.

(c) *If $F = \mathbf{R}$ and H is indefinite, then $\mathcal{U}(H, \mathbf{R})$ has four connected components which can be described as follows. We can assume $H = I_p \oplus -I_q$, where $p, q > 0$.*

Then, for every choice of signs $\delta_1 = \pm 1$, $\delta_2 = \pm 1$, a connected component of $\mathcal{U}(H, \mathbf{R})$ is given by

$$\mathcal{U}(H; \delta_1, \delta_2) = \left\{ V = \begin{bmatrix} V_1 & V_2 \\ V_3 & V_4 \end{bmatrix} \in \mathcal{U}(H; \mathbf{R}) |\ \delta_1 \det V_1 > 0,\ \delta_2 \det V_4 > 0 \right\},$$

where V_1 is a $p \times p$ matrix and V_4 is a $q \times q$ matrix. In particular,

$$\{X \in \mathcal{U}(H; \mathbf{R}) |\ \det X = 1\} = \mathcal{U}(H; 1, 1) \cup \mathcal{U}(H; -1, -1), \tag{3.1}$$

and this set consists of two connected components.

In all cases, each connected component of $\mathcal{U}(H; F)$ is arcwise connected.

Proof. This result is known; for the proof of (a) and (b) see Lemma I.3.8 and Theorem I.5.8, respectively, in [GLR].

For completeness, we provide a proof of (c). Let $V = \begin{bmatrix} V_1 & V_2 \\ V_3 & V_4 \end{bmatrix}$ belong to $\mathcal{U}(H; \mathbf{R})$. Then the equation $V^T H V = H = I_p \oplus -I_q$ gives

$$V_1^T V_1 = I + V_3^T V_3, \quad V_4^T V_4 = I + V_2^T V_2, \quad V_2^T V_1 = V_4^T V_3. \tag{3.2}$$

It follows that $|\det V_1| \geq 1$, $|\det V_4| \geq 1$, and therefore the H-unitary matrices (over $\mathbf{R}$) $V = \begin{bmatrix} V_1 & V_2 \\ V_3 & V_4 \end{bmatrix}$ and $W = \begin{bmatrix} W_1 & W_2 \\ W_3 & W_4 \end{bmatrix}$ (here W_1 is $p \times p$ and W_4 is $q \times q$) belong to different connected components in $\mathcal{U}(H; \mathbf{R})$, provided at least one of the inequalities $\det V_1 \cdot \det W_1 < 0$, $\det V_4 \cdot \det W_4 < 0$ is valid. It remains to show that if $\det V_1 \cdot \det W_1 > 0$ and $\det V_4 \cdot \det W_4 > 0$, then V and W belong to the same connected component in $\mathcal{U}(H, \mathbf{R})$. It suffices to show that if $\det V_1 > 0$, $\det V_4 > 0$, then V can be continuously connected to I in $\mathcal{U}(H; \mathbf{R})$. As V is H-unitary, V^T is H-unitary as well (indeed, $V^T H V = H$ implies $V^{-1} = H^{-1} V^T H = H V^T H$, and therefore $V H V^T H = I$, or $V H V^T = H$). Thus, we also have

$$I + V_3 V_3^T = V_4 V_4^T. \tag{3.3}$$

Observe from (3.2) and (3.3) that $V_1 (I + V_3^T V_3)^{-\frac{1}{2}}$ and $V_4^T (I + V_3 V_3^T)^{-\frac{1}{2}}$ are real and unitary (with respect to I). Moreover, they both have determinant 1, as $\det V_1 > 0$ and $\det V_4 > 0$. So, by part (b), there is a continuous family of unitary matrices $U_1(t)$, $U_4(t)$ for $t \in [0, 1]$ such that

$$U_1(0) = I, \qquad U_4(0) = I,$$
$$U_1(1) = V_1 (I + V_3^T V_3)^{-\frac{1}{2}}, \qquad U_4(1) = V_4^T (I + V_3 V_3^T)^{-\frac{1}{2}}.$$

Let

$$V_1(t) = U_1(t)(I + t^2 V_3^T V_3)^{\frac{1}{2}}, \qquad V_4(t) = (I + t^2 V_3 V_3^T)^{\frac{1}{2}} U_4(t)^T,$$
$$V_3(t) = t V_3, \qquad V_2(t) = t V_1(t)^{-T} V_3^T V_4(t),$$

and

$$V(t) = \begin{bmatrix} V_1(t) & V_2(t) \\ V_3(t) & V_4(t) \end{bmatrix}.$$

Then $V(0) = I$ and $V(1) = V$, and one easily verifies that $V(t)$ is H-unitary for all $t \in [0, 1]$. □

A basis independent description of the connected components of $\mathcal{U}(H; \mathbf{R})$, where H is indefinite, runs as follows. Let $\mathcal{M}_+$ and $\mathcal{M}_-$ be subspaces in $\mathbf{R}^n$ which are H-orthogonal complements of each other and such that $\mathcal{M}_+$ is H-positive and $\mathcal{M}_-$ is H-negative. Denote by P_+ (resp., P_-) the projector onto $\mathcal{M}_+$ (resp., $\mathcal{M}_-$) along $\mathcal{M}_-$ (resp., $\mathcal{M}_+$). Then $X \in \mathcal{U}(H; \delta_1, \delta_2)$ if and only if

$$\delta_1 \det(P_+ X|M_+) > 0, \qquad \delta_2 \det(P_- X|\mathcal{M}_-) > 0. \tag{3.4}$$

The proof of this statement is analogous to the proof of Theorem 3.1 (part (c)) and therefore is omitted.

Observe that the inequalities (3.4) are independent of the choice of the pair of subspaces $\mathcal{M}_+$, $\mathcal{M}_-$ with the above properties.

THEOREM 3.2. *For any real invertible matrix S and any $X \in \mathcal{U}(H; \delta_1, \delta_2)$ the matrix $S^{-1}XS$ belongs to the connected component $\mathcal{U}(S^*HS; \delta_1, \delta_2)$ determined by the same δ_1, δ_2.*

Proof. The proof follows easily from the description of $\mathcal{U}(H; \delta_1, \delta_2)$ given by formula (3.4). Indeed, assume that $X \in \mathcal{U}(H; \delta_1, \delta_2)$. Choose a pair of subspaces $\mathcal{M}_+$ and $\mathcal{M}_-$ that are H-orthogonal complements to each other and such that $\mathcal{M}_+$ (resp., $\mathcal{M}_-$) is H-positive (resp., H-negative). Then $S^{-1}\mathcal{M}_+$ and $S^{-1}\mathcal{M}_-$ are S^*HS-orthogonal complements to each other and $S^{-1}\mathcal{M}_+$ (resp., $S^{-1}\mathcal{M}_-$) is S^*HS-positive (resp., S^*HS-negative). We conclude the proof by applying the formula (3.4) with X, P_+, P_- replaced by $S^{-1}XS$, $S^{-1}P_+S$, $S^{-1}P_-S$, respectively, and with $\mathcal{M}_\pm$ replaced by $S^{-1}\mathcal{M}_\pm$. □

4. Witt's theorem for real skew-symmetric scalar products. Let $F = \mathbf{R}$ and let K be a real invertible skew-Hermitian $n \times n$ matrix (in particular, n is even). Define the skew-symmetric scalar product $\{\cdot, \cdot\}$ on $\mathbf{R}^n$ by

$$\{x, y\} = \langle Kx, y\rangle.$$

If A is an $n \times n$ matrix, its K-adjoint $A^{\{*\}}$ is defined by the identity $\{Ax, y\} = \{x, A^{\{*\}}y\}$, where $x, y \in \mathbf{R}^n$. It is easy to see that $A^{\{*\}} = K^{-1}A^*K$. A matrix A is called *K-self-adjoint* if $A^{\{*\}} = A$, and it is called *K-skew-self-adjoint* if $A^{\{*\}} = -A$. A *K-unitary* matrix A is defined by the property that it preserves the skew-symmetric scalar product, i.e., if, for any two vectors $x, y \in \mathbf{R}^n$, $\{Ax, Ay\} = \{x, y\}$. It is easy to verify that A is K-self-adjoint if and only if $KA = A^*K$, is K-skew-self-adjoint if and only if $KA = -A^*K$, and is K-unitary if and only if it is nonsingular and $K^{-1}A^*KA = I$.

Example 4.1. Consider the skew-Hermitian matrix

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

We have $H^* = H^{-1} = -H$. Moreover, $X^{\{*\}} = H^{-1}X^*H$ is the cofactor matrix of X, so that $X^{\{*\}}X = (\det X)\, I$. Hence A is H-self-adjoint if and only if $A = cI$ for some $c \in \mathbf{R}$, and A is H-skew-self-adjoint (i.e., $HA = -A^*H$) if and only if $\operatorname{Tr} A = 0$. Furthermore, U is H-unitary (i.e., $U^*HU = H$) if and only if $\det U = +1$. □

LEMMA 4.1. *Let $\{.,.\}$ be a skew-symmetric scalar product on $\mathbf{R}^n$ defined by the real invertible skew-symmetric $n \times n$ matrix K and let V be an m-dimensional subspace*

of $\mathbf{R}^n$. *Let the defect of the restriction of* $\{.,.\}$ *to* V *be* m_0 (*so that the rank of the above restriction is* $m - m_0$). *Then*

(a) *There exists a basis*

$$\{e_1, \dots, e_{m_0}, e_{m_0+1}, \dots, e_{\frac{m+m_0}{2}}, f_{m_0+1}, \dots, f_{\frac{m+m_0}{2}}\} \tag{4.1}$$

of V *such that*

$$\{e_k, f_k\} = -\{f_k, e_k\} = 1, \qquad k = m_0 + 1, m_0 + 2, \dots, \frac{m + m_0}{2}, \tag{4.2}$$

while the scalar product of any other two vectors in (4.1) *is zero.*

(b) *There exist vectors*

$$\{f_1, f_2, \dots, f_{m_0}, e_{\frac{m+m_0}{2}+1}, e_{\frac{m+m_0}{2}+2}, \dots, e_{\frac{n}{2}}, f_{\frac{m+m_0}{2}+1}, f_{\frac{m+m_0}{2}+2}, \dots, f_{\frac{n}{2}}\} \tag{4.3}$$

such that the union of the sets (4.1) *and* (4.3) *is a canonical basis for* $\mathbf{R}^n$*; i.e.;*

$$\{e_k, f_k\} = -\{f_k, e_k\} = 1, \ \ k = 1, 2, \dots, \frac{n}{2}, \tag{4.4}$$

while the scalar product of any other two vectors from the union of (4.1) *and* (4.2) *is zero.*

Proof. This is an elementary exercise in linear algebra. Namely, if $m = m_0$ then V is isotropic and any basis of V does the job. If $m > m_0$ there exist vectors e_{m_0+1} and f_{m_0+1} such that $\{e_{m_0+1}, f_{m_0+1}\} = 1$. If $m - m_0 = 2$ then the orthogonal companion V_1 of the subspace $\operatorname{span}\{e_{m_0+1}, f_{m_0+1}\}$ in V is isotropic and any basis of V_1 appended to vectors e_{m_0+1}, f_{m_0+1} produces a desired basis. If $m - m_0 > 2$ then V_1 is not isotropic and we can find vectors $e_{m_0+2}, f_{m_0+2} \in V_1$ such that $\{e_{m_0+2}, f_{m_0+2}\} = 1$. Continuing this process we will find a desired basis of V. This proves (a). To prove (b) we first introduce the $(n - m + m_0)$-dimensional subspace W of $\mathbf{R}^n$, which is K-orthogonal to the subspace

$$\operatorname{span}\{e_{m_0+1}, f_{m_0+1}, e_{m_0+2}, f_{m_0+2}, \dots, e_{\frac{m+m_0}{2}}, f_{\frac{m+m_0}{2}}\}.$$

Obviously, W is nondegenerate and $e_1, e_2, \dots, e_{m_0} \in W$. Since W is nondegenerate, there exists a vector $f_1 \in W$ such that $\{e_1, f_1\} = 1$ and $\{e_k, f_1\} = 0$ for $k = 2, 3, \dots, m_0$. Let W_1 be the K-orthogonal complement of $\operatorname{span}\{e_1, f_1\}$ in W. If $m_0 = 1$ then any basis of W_1 appended to vectors e_1 and f_1 already found will produce a desired basis. If $m_0 > 1$ then $e_2 \in W_1$ and we can find a vector $f_2 \in W_1$ such that $\{e_2, f_2\} = 1$ and $\{e_k, f_2\} = 0$ for $k = 1, 3, 4, \dots, m_0$. Continuing this process we will finally find a basis of $\mathbf{R}^n$ that satisfies all the requirements of (b). ∎

THEOREM 4.2. *Let* $\{\cdot,\cdot\}_1$ *and* $\{\cdot,\cdot\}_2$ *be two skew-symmetric scalar products on* $\mathbf{R}^n$ *defined by the skew-symmetric* $n \times n$ *matrices* K_1 *and* K_2*, respectively:*

$$\{x, y\}_1 = \langle K_1 x, y\rangle, \ \ \{x, y\}_2 = \langle K_2 x, y\rangle, \ \ x, y \in \mathbf{R}^n.$$

Let $U_0 : V_1 \to V_2$*, where* V_1 *and* V_2 *are subspaces in* $\mathbf{R}^n$*, be a nonsingular linear transformation that preserves the scalar products; namely,*

$$\{U_0 x, U_0 y\}_2 = \{x, y\}_1$$

for every $x, y \in V_1$*. Then there exists a linear transformation* $U : \mathbf{R}^n \to \mathbf{R}^n$ *such that*

$$\{Ux, Uy\}_2 = \{x, y\}_1$$

for every $x, y \in V_1$ *and*

$$Ux = U_0 x$$

for every $x \in V_1$.

Proof. Let the vectors

$$\{e_1, \dots, e_{m_0}, e_{m_0+1}, \dots, e_{\frac{m+m_0}{2}}, f_{m_0+1}, \dots, f_{\frac{m+m_0}{2}}\} \tag{4.5}$$

be as in (a) of Lemma 4.1 and let

$$g_t = U_0 e_t, \quad h_s = U_0 f_s, \quad t = 1, 2, \dots, \frac{m+m_0}{2}, \ s = m_0+1, m_0+2, \dots, \frac{m+m_0}{2}. \tag{4.6}$$

Next, let the vectors

$$f_1, f_2, \dots, f_{m_0}, e_{\frac{m+m_0}{2}+1}, e_{\frac{m+m_0}{2}+2}, \dots, e_{\frac{n}{2}}, f_{\frac{m+m_0}{2}+1}, f_{\frac{m+m_0}{2}+2}, \dots, f_{\frac{n}{2}} \tag{4.7}$$

be as in (b) of Lemma 4.1; i.e., combined with the vectors (4.1) they produce a canonical basis

$$\{e_1, e_2, \dots, e_{m_0}, e_{m_0+1}, e_{m_0+2}, \dots, e_{\frac{m+m_0}{2}}, f_{m_0+1}, f_{m_0+2}, \dots, f_{\frac{m+m_0}{2}},$$

$$f_1, f_2, \dots, f_{m_0}, e_{\frac{m+m_0}{2}+1}, e_{\frac{m+m_0}{2}+2}, \dots, e_{\frac{n}{2}}, f_{\frac{m+m_0}{2}+1}, f_{\frac{m+m_0}{2}+2}, \dots, f_{\frac{n}{2}}\} \tag{4.8}$$

of $\mathbf{R}^n$,

$$\{e_s, f_s\}_1 = -\{f_s, e_s\}_1 = 1, \qquad s = 1, 2, \dots, \frac{n}{2}. \tag{4.9}$$

The remaining scalar products of the basis are zero. Similarly, let the vectors

$$h_1, h_2, \dots, h_{m_0}, g_{m+1}, g_{m+2}, \dots, g_{\frac{n}{2}}, h_{m+1,}, h_{m+2}, \dots, h_{\frac{n}{2}} \tag{4.10}$$

be as in (b) of Lemma 4.1; i.e., combined with the vectors (4.6) they produce a canonical basis

$$\{g_1, g_2, \dots, g_{m_0}, g_{m_0+1}, g_{m_0+2}, \dots, g_{\frac{m+m_0}{2}}, h_{m_0+1}, h_{m_0+2}, \dots, h_{\frac{m+m_0}{2}},$$

$$h_1, h_2, \dots, h_{m_0}, g_{\frac{m+m_0}{2}+1}, g_{\frac{m+m_0}{2}+2}, \dots, g_{\frac{n}{2}}, h_{\frac{m+m_0}{2}+1}, h_{\frac{m+m_0}{2}+2}, \dots, h_{\frac{n}{2}}\} \tag{4.11}$$

of $\mathbf{R}^n$,

$$\{g_s, h_s\}_2 = -\{h_s, g_s\}_2 = 1, \qquad s = 1, 2, \dots, \frac{n}{2}. \tag{4.12}$$

The remaining scalar products of the basis are zero. Define the linear transformation U as follows:

$$Ue_s = g_s, \quad Uf_s = h_s, \quad s = 1, 2, \dots, \frac{n}{2}. \tag{4.13}$$

It is easy to see that the matrix defined by (4.13) satisfies all the conditions of the theorem. ∎

We will use the bases (4.8) and (4.11) constructed in the proof of Theorem 4.2. With respect to $\{.,.\}_1$, the basis (4.8) has the skew-symmetric Gramian matrix

$$K = \begin{bmatrix} 0 & 0 & -I & 0 \\ 0 & J_1 & 0 & 0 \\ I & 0 & 0 & 0 \\ 0 & 0 & 0 & J_2 \end{bmatrix}. \tag{4.14}$$

Here I is the $m_0 \times m_0$ identity matrix, J_1 is an $(m-m_0)\times(m-m_0)$ matrix of the form $J_1 = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}$, and J_2 is an $(n-m-m_0)\times(n-m-m_0)$ matrix of the same form as J_1.

As in previous sections, any linear transformation (or its matrix representation with respect to fixed bases) U from Theorem 4.2 will be called a *Witt extension* of U_0. All the Witt extensions of a given U_0 are described by the following theorem (we represent the Witt extensions as linear transformations $\mathbf{R}^n \to \mathbf{R}^n$ with respect to the bases (4.8) and (4.11) above).

THEOREM 4.3 (extended Witt's theorem for a skew-symmetric scalar product). *If a matrix $\tilde{U}$ is a Witt extension of the matrix U_0, then there exist a J_2-unitary matrix P_1 (of order $n-m-m_0$), a real $(n-m-m_0)\times m_0$ matrix P_2, and a real symmetric $m_0 \times m_0$ matrix P_3 (i.e., $P_3^* = P_3$) such that the matrix of $\tilde{U}$ has the form*

$$\tilde{U} = \begin{bmatrix} I & 0 & -\frac{1}{2}P_2^* J_2 P_2 + P_3 & -P_2^* J_2 P_1 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & P_2 & P_1 \end{bmatrix}. \tag{4.15}$$

Conversely, if P_1 is an arbitrary J_2-unitary matrix, P_2 is an arbitrary real $(n-m-m_0)\times m_0$ matrix, and P_3 is an arbitrary real symmetric $m_0 \times m_0$ matrix, then the matrix $\tilde{U}$ defined by (4.15) *is a Witt extension of U_0.*

Proof. The proof is similar to that of Theorem 2.3. Namely, any extension $\tilde{U}$ of U_0 in the bases (4.8), (4.11) has the matrix

$$\tilde{U} = \begin{bmatrix} I & 0 & A_1 & A_2 \\ 0 & I & A_3 & A_4 \\ 0 & 0 & A_5 & A_6 \\ 0 & 0 & A_7 & A_8 \end{bmatrix}. \tag{4.16}$$

The necessary and sufficient condition for the matrix $\tilde{U}$ to be K_1-K_2-unitary is the identity $K_1^{-1}\tilde{U}^* K_2 \tilde{U} = I$. Taking into account (4.14), (4.16), and the facts that $K_1 = K_2 = K$ and that $K^{-1} = -K$, we can rewrite the last relation in block form as

$$\begin{bmatrix} A_5^* & A_3^* J_1 & u_{13} & u_{14} \\ 0 & I & A_3 & A_4 \\ 0 & 0 & A_5 & A_6 \\ -J_2 A_6^* & -J_2 A_4^* J_1 & u_{43} & u_{44} \end{bmatrix} = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix}, \tag{4.17}$$

where

$$u_{13} = -A_1^* A_5 + A_3^* J_1 A_3 + A_5^* A_1 + A_7^* J_2 A_7,$$

$$u_{14} = -A_1^* A_6 + A_3^* J_1 A_4 + A_5^* A_2 + A_7^* J_2 A_8,$$

$$u_{43} = J_2 A_2^* A_5 - J_2 A_4^* J_1 A_3 - J_2 A_6^* A_1 - J_2 A_8^* J_2 A_7,$$

$$u_{44} = J_2 A_2^* A_6 - J_2 A_4^* J_1 A_4 - J_2 A_6^* A_2 - J_2 A_8^* J_2 A_8.$$

Equating the corresponding blocks in (4.17), we easily derive the statement of the theorem. The only appropriate clarification to make is the following. After we establish that $A_3 = A_4 = A_6 = 0$ and that $A_5 = I$ we can rewrite the equation $u_{13} = 0$ as

$$A_1 - A_1^* + A_7^* J_2 A_7 = 0. \tag{4.18}$$

Having represented the matrix A_1 as a sum of symmetric and skew-symmetric matrices, we get $A_1 = A_+ + A_-$, where $A_+^* = A_+$ and $A_-^* = -A_-$. Substituting $A_+ + A_-$ for A_1 and $A_+ - A_-$ for A_1^* into (4.18) we conclude that

$$A_- = -\frac{1}{2} A_7^* J_2 A_7$$

and that, for an arbitrary self-adjoint matrix P_3, the matrix $A_1 = P_3 - \frac{1}{2} A_7^* J_2 A_7$ satisfies the equation (4.18). □

The formula (2.9) for the inverse of $\tilde{U}$ is valid here as well.

Note that a Witt extension $\tilde{U}$ has the form (4.17) with respect to different bases in domain and image space, namely, with respect to the basis (4.8) in the domain space and the basis (4.11) consisting of the vectors Ue, in the image space. Keeping this in mind, we can easily reformulate Theorem 4.2 in the following way with respect to one basis (4.8) (the same for both the domain and the range of $\tilde{U}$) and obtain a statement similar to Theorem 2.4.

THEOREM 4.4 (extended Witt's theorem, second version). *Let U be a fixed Witt extension of U_0 as constructed in Theorem* 4.1. *Then any Witt extension $\tilde{U}$ of U_0 is given by $\tilde{U} = UM$, where M has the form of the right-hand side of* (4.16) *with respect to the basis* (4.8).

Proof. Observe that U maps the elements of the basis (4.8) into the corresponding elements of the basis (4.11) and that (4.15) is the matrix representation of U with respect to the basis (4.8) in the domain space and the basis (4.11) in the image space. □

The set of all Witt extensions of an isometry between two real skew-symmetric scalar product spaces is described as follows.

THEOREM 4.5. *Let H_1, H_2, V_1, and U_0 be as in Theorem* 4.2. *Then the set $W(U_0)$ of all Witt extensions of U_0 is connected and can be parametrized by $\frac{1}{2}(n-m)(n-m+1)$ real variables. More precisely, let*

$$m_0 = \delta \left[z_j^T (iH_1) z_k \right]_{j,k=1}^{m}$$

for some (every) basis $\{z_1, \dots, z_m\}$ in V_1; in other words, m_0 is the defect of the restriction of H_1 to V_1. Then $W(U_0)$ is diffeomorphic (as a real analytic manifold) to

$$SU\left(\frac{n-m-m_0}{2}\right) \times T \times \mathbf{R}^u,$$

where

$$u = \frac{n-m-m_0}{2}\left(\frac{n-m+3m_0}{2}+1\right) + \frac{m_0(m_0+1)}{2}.$$

The proof is obtained by combining Theorem 4.3 and the parametrization of the group of all real matrices that are orthogonal with respect to a skew-symmetric scalar

product (see Theorem II.1.7 in [GLR]). Observe that this group is connected (see the same theorem in [GLR]). Also observe that the set of (real) J_2-unitary matrices is diffeomorphic (as a real analytic manifold) to $\mathbf{R}^{k(k+1)} \times SU(k) \times T$, where $k = \frac{n-m-m_0}{2}$, and hence can be described by $2k^2 + k$ real parameters; a detailed proof is found in section II.1.5 of [GLR].

As in Theorems 2.5 and 2.6, the number of independent parameters that describe the set of Witt extensions in Theorem 4.5 depends only on n and m and does not depend on m_0.

5. Polar decompositions. Let $F = \mathbf{C}$ or $F = \mathbf{R}$, and let H be an invertible Hermitian $n \times n$ matrix over F. A factorization $X = UA$ will be called a *semidefinite H-polar decomposition* if U is H-unitary, A is H-nonnegative, and both U and A are over F. Recall that an $n \times n$ matrix A is said to be H-nonnegative if HA is positive semidefinite Hermitian.

More general classes and concepts of polar decompositions in indefinite scalar product spaces are studied in [BMRRR1]. If H is positive definite, then the concept of semidefinite H-polar decomposition reduces to the well-known and widely used notion of polar decompositions for real and complex matrices. For an indefinite H, polar decompositions have been studied in [P1, P2, AI1, AI2, BMRRR2] in connection with Potapov's theory of H-nonexpansive operators, in [KS1, KS2] in connection with plus operators, and in [BR] in connection with H-unitary equivalence. Such polar decompositions play an important role in certain applications in linear optics [M, MH, BMRRR2]. A general approach to polar decompositions is developed in [K]. Other variants of factorizations of matrices of the polar decomposition type have also been studied extensively in the literature; see, e.g., [HM1, HM2, CH].

In this section we characterize the matrices X which admit semidefinite H-polar decompositions (note that in contrast to the standard polar decompositions not every real or complex matrix admits H-polar decompositions if H is indefinite; see [BMRRR1] for examples). Furthermore, in the case when semidefinite H-polar decompositions exist, we provide a full description of the H-nonnegative and H-unitary factors.

We start by recalling the canonical forms of H-self-adjoint matrices (more precisely, of the pairs $\{A, H\}$, where A is H-self-adjoint). We denote by $J_k(\lambda)$ the $k \times k$ upper triangular Jordan block with $\lambda \in \mathbf{C}$ on the main diagonal and by $J_k(\lambda \pm i\mu)$ the $k \times k$ almost upper triangular real Jordan block with eigenvalues $\lambda \pm i\mu$ (here λ, μ are real and $\mu > 0$; k is necessarily even). We also use the notation $Q_m = [\delta_{i+j,m+1}]_{i,j=1}^m$ for the $m \times m$ matrix with ones on the southwest–northeast diagonal and zeros elsewhere.

THEOREM 5.1. *Let H be an $n \times n$ invertible Hermitian matrix (over F), and let $A \in F^{n\times n}$ be H-self-adjoint. Then there exists an invertible S over F such that $S^{-1}AS$ and S^*HS have the form*

$$S^{-1}AS = J_{k_1}(\lambda_1) \oplus \cdots \oplus J_{k_\alpha}(\lambda_\alpha) \oplus \left[J_{k_{\alpha+1}}(\lambda_{\alpha+1}) \oplus J_{k_{\alpha+1}}(\bar{\lambda}_{\alpha+1})\right] \oplus \cdots \oplus \left[J_{k_\beta}(\lambda_\beta) \oplus J_{k_\beta}(\bar{\lambda}_\beta)\right] \tag{5.1}$$

if $F = \mathbf{C}$, where $\lambda_1, \ldots, \lambda_\alpha$ are real and $\lambda_{\alpha+1}, \ldots, \lambda_\beta$ are nonreal with positive imaginary parts;

$$S^{-1}AS = J_{k_1}(\lambda_1) \oplus \cdots \oplus J_{k_\alpha}(\lambda_\alpha) \oplus J_{2k_{\alpha+1}}(\lambda_{\alpha+1} \pm i\mu_{\alpha+1}) \oplus \cdots \oplus J_{2k_\beta}(\lambda_\beta \pm i\mu_\beta) \tag{5.2}$$

if $F = \mathbf{R}$, where $\lambda_1, \dots, \lambda_\beta$ are real and $\mu_{\alpha+1}, \dots, \mu_\beta$ are positive;

$$S^*HS = \epsilon_1 Q_{k_1} \oplus \cdots \oplus \epsilon_\alpha Q_{k_\alpha} \oplus Q_{2k_{\alpha+1}} \oplus \cdots \oplus Q_{2k_\beta} \tag{5.3}$$

for both cases ($F = \mathbf{R}$ or $F = \mathbf{C}$), where $\epsilon_1, \dots, \epsilon_\alpha$ are ± 1. For a given pair $\{A, H\}$, where A is H-self-adjoint, the canonical form (5.1), (5.2), (5.3) *is unique up to permutation of orthogonal components in* (5.3), *and the same simultaneous permutation of the corresponding blocks in* (5.1) *or* (5.2), *as the case may be.*

Theorem 5.1 is well known and goes back to Weierstrass and Kronecker. A complete proof of this theorem can be found in many sources; see, e.g., [GLR, T].

The signs ϵ_j in (5.3) form the *sign characteristic* of the pair $\{A, H\}$. Thus, the sign characteristic consists of signs $+1$ or -1 attached to every partial multiplicity (= size of a Jordan block in the Jordan form) of A corresponding to a real eigenvalue.

An existence result concerning general classes of polar decompositions with respect to indefinite scalar products was proved in [BMRRR1, Theorem 4.1]. In particular, this theorem contains the following statement.

PROPOSITION 5.2. *An $n \times n$ matrix X (over F) admits a semidefinite H-polar decomposition if and only if $X^{[*]}X = A^2$ for some H-nonnegative matrix A such that* $\operatorname{Ker} A = \operatorname{Ker} X$*; moreover, for any such A there is an H-unitary U such that $X = UA$.*

This existence result can be given a much more tractable formulation.

THEOREM 5.3 ($F = \mathbf{C}$ *or* $F = \mathbf{R}$). *An $n \times n$ matrix X admits a semidefinite H-polar decomposition if and only if $X^{[*]}X$ has eigenvalues only in $\{\lambda \in \mathbf{R} |\ \lambda \geq 0\}$ and is diagonalizable and moreover, if* $\operatorname{Ker} X$ *contains a k-dimensional H-nonpositive subspace, where k is the number of negative signs in the sign characteristic of $\{X^{[*]}X, H\}$ corresponding to the zero eigenvalue, and* $\operatorname{Ker} X$ *contains a p-dimensional H-nonnegative subspace, where p is the number of positive signs of H corresponding to the zero eigenvalue of $X^{[*]}X$. Moreover, A can be chosen as to satisfy* $\operatorname{Ker}(A^2) = \operatorname{Ker} A$ *if and only if the subspace* $\operatorname{Ker} X^{[*]}X = \operatorname{Ker} X$ *is H-nondegenerate.*

Proof. Suppose X admits a semidefinite H-polar decomposition $X = UA$. Then

$$X^{[*]}X = A^2. \tag{5.4}$$

Since A is H-nonnegative, the canonical form (Theorem 5.1) for $\{A, H\}$ implies that there is an invertible matrix S (over F) such that

$$S^{-1}AS = \operatorname{diag}(\lambda_i)_{i=1}^{\nu_1} \oplus 0_{\nu_2} \oplus \operatorname{diag}\left(\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\right)_{i=1}^{\nu_3} \oplus \operatorname{diag}(\mu_i)_{i=1}^{\nu_4}, \tag{5.5}$$

where λ_i are negative, μ_i are positive, and

$$S^*HS = -I_{\nu_1} \oplus \operatorname{diag}(\epsilon_i)_{i=1}^{\nu_2} \oplus \operatorname{diag}\left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\right)_{i=1}^{\nu_3} \oplus I_{\nu_4}, \tag{5.6}$$

where $\epsilon_i = \pm 1$. Then

$$S^{-1}A^2S = \operatorname{diag}(\lambda_i^2)_{i=1}^{\nu_1} \oplus 0_{(\nu_2+2\nu_3)} \oplus \operatorname{diag}(\mu_i^2)_{i=1}^{\nu_4},$$

and thus A^2 is diagonalizable with nonnegative eigenvalues. In view of (5.4) the same thing is true of $X^{[*]}X$. Now we show that $\operatorname{Ker} X = \operatorname{Ker} A$ contains a k-dimensional

H-nonpositive subspace and a p-dimensional H-nonnegative subspace. This follows easily from (5.5), (5.6), as in the notation introduced there:

$$p = \nu_3 + \#\{\epsilon_i |\ \epsilon_i = +1,\ i = 1, \dots, \nu_2\},$$
$$k = \nu_3 + \#\{\epsilon_i |\ \epsilon_i = -1,\ i = 1, \dots, \nu_2\}.$$

To prove the converse part we will need the following lemma (its proof can be found in [BMRRR1]).

LEMMA 5.4. *Let $H = H^*$ be an invertible $n \times n$ matrix, and let X be an $n \times n$ matrix. Let S be an invertible $n \times n$ matrix such that*

$$S^{-1} X^{[*]} X S = \operatorname{diag}(Z_i)_{i=1}^{\nu}, \quad S^* H S = \operatorname{diag}(H_i)_{i=1}^{\nu},$$

with $\sigma(Z_i) \cap \sigma(Z_j) = \emptyset$ for $i \neq j$. Then there exists an H-self-adjoint, respectively, H-nonnegative, matrix A such that $X^{[*]}X = A^2$ if and only if for each i there exists an H_i-self-adjoint, respectively, H_i-nonnegative, matrix A_i such that $Z_i = A_i^2$.

To prove the "if" part of Theorem 5.3, we now only have to consider the case where $X^{[*]}X$ has a single eigenvalue, $\sigma(X^{[*]}X) = \{\lambda\}$. The cases $\lambda > 0$ and $\lambda = 0$ will be considered separately.

Suppose $X^{[*]}X$ is diagonalizable and $\sigma(X^{[*]}X) = \{\lambda\},\ \lambda > 0$. Let S be an invertible matrix such that

$$S^{-1} X^{[*]} X S = \begin{bmatrix} \lambda I_{n_1} & 0 \\ 0 & \lambda I_{n_2} \end{bmatrix}, \qquad S^* H S = \begin{bmatrix} I_{n_1} & 0 \\ 0 & -I_{n_2} \end{bmatrix}.$$

The existence of S is guaranteed; in fact, one brings the pair $\{X^{[*]}X, H\}$ to the canonical form in this way (Theorem 5.1). Let

$$A = S \begin{bmatrix} \sqrt{\lambda}\, I_{n_1} & 0 \\ 0 & -\sqrt{\lambda}\, I_{n_2} \end{bmatrix} S^{-1}.$$

Then A is H-nonnegative and $A^2 = X^{[*]}X$.

Finally, assume $X^{[*]}X$ is diagonalizable, $\sigma(X^{[*]}X) = \{0\}$ (then $X^{[*]}X = 0$), and $\operatorname{Ker} X$ contains a k-dimensional H-nonpositive subspace and a p-dimensional H-nonnegative subspace. It is easy to see that $k + p = n$ in this case, so $\operatorname{Ker} X$ contains a maximal H-nonpositive subspace and a maximal H-nonnegative subspace. For the sake of convenience write $M = \operatorname{Ker} X$. Put $N = M \cap (HM)^{\perp}$, and let M_1 be such that $M = N \oplus M_1$, where this direct sum is orthogonal. This direct sum is also H-orthogonal. Select a basis $f_1, \dots, f_{\nu^0}$ in N and a basis $e_1, \dots, e_{\nu^+}, e_{\nu^+ + 1}, \dots, e_{\nu^+ + \nu^-}$ in M_1 such that

$$\langle He_i, e_j \rangle = 0 \quad \text{for} \quad i \neq j,$$

$$\langle He_i, e_i \rangle = 1 \quad \text{if} \quad i \leq \nu^+, \qquad \langle He_i, e_i \rangle = -1 \quad \text{if} \quad i > \nu^+.$$

We shall construct a subspace K such that $M \oplus K = F^n$ and $(HK)^{\perp} = K \oplus M_1$.

We shall construct an H-nonnegative matrix A such that $A^2 = 0$ and $\operatorname{Ker} A = \operatorname{Ker} X$. The matrix A will be constructed so that N coincides with the linear span of eigenvectors of A corresponding to Jordan blocks of length 2, while $N \oplus K$ is spanned by the eigenvectors, as well as by the generalized eigenvectors of A.

As M contains a maximal H-nonnegative and a maximal H-nonpositive subspace we have

$$\nu^+ + \nu^0 = k,\ \nu^- + \nu^0 = p,$$

and therefore $\dim M = \nu^0 + \nu^+ + \nu^- = k + p - \nu^0 = n - \nu^0$. Consider $(HM_1)^\perp$. The dimension of this subspace is $n - \nu^+ - \nu^- = k + p - \nu^+ - \nu^- = 2\nu^0$; moreover, $(HM_1)^\perp$ contains N. Take any subspace K' such that $(HM_1)^\perp = N \oplus K'$. Then K' is a direct complement of M. Indeed, as $\nu^0 = \dim N = \operatorname{codim} M$ we have $(HM)^\perp = N, (HN)^\perp = M$. Therefore,

$$N = (HM)^\perp = (H(N \oplus M_1))^\perp = (HN)^\perp \cap (HM_1)^\perp = M \cap (HM_1)^\perp.$$

So $K' \cap M = (0)$. Also, $\dim K' = \nu^0$. Take vectors $g'_1, \dots, g'_{\nu^0}$ in K' such that $\langle Hf_i, g'_j \rangle = \delta_{ij}$ for $i, j = 1, \dots, \nu^0$. Construct

$$g_i = g'_i - \frac{1}{2} \sum_{\nu=1}^{\nu^0} \langle Hg'_i, g'_\nu \rangle f_\nu, \quad i = 1, \dots, \nu^0,$$

and let $K = \operatorname{span}\{g_1, \dots, g_{\nu^0}\}$. Then

$$\langle Hg_i, g_j \rangle = 0 \quad \text{for all } i, j,$$

$$\langle Hf_i, g_j \rangle = \delta_{ij} \text{ for all } i, j,$$

and $K \subset N \oplus K' = (HM_1)^\perp$. By construction, K is H-neutral, so $(HK)^\perp = K \oplus M_1$. Consider the vectors

$$e_1, \dots, e_{\nu^+}, e_{\nu^+ + 1}, \dots, e_{\nu^+ + \nu^-}, f_1, g_1, f_2, g_2, \dots, f_{\nu^0}, g_{\nu^0}$$

as a basis for F^n, and let S be the matrix with these basis vectors as its columns in the order in which they appear here. Then

$$S^* H S = I_{\nu^+} \oplus -I_{\nu^-} \oplus \operatorname{diag} \left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \right)_{i=1}^{\nu^0}.$$

Construct A as follows:

$$S^{-1} A S = 0_{(\nu^+ + \nu^-)} \oplus \operatorname{diag} \left(\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \right)_{i=1}^{\nu^0}.$$

Then, A is H-nonnegative, $A^2 = 0$, and

$$\operatorname{Ker} A = \operatorname{span}\{e_1, \dots, e_{\nu^+}, e_{\nu^+ + 1}, \dots, e_{\nu^+ + \nu^-}, f_1, \dots, f_{\nu^0}\} = \operatorname{Ker} X.$$

By Proposition 5.2 X admits a semidefinite H-polar decomposition.

The statement on choosing A to satisfy $\operatorname{Ker}(A^2) = \operatorname{Ker} A$ is clear because it is equivalent to the nondegeneracy of $\operatorname{Ker} A$. Further, when constructing such A as above, one has $\nu^0 = 0$, which implies $\operatorname{Ker}(A^2) = \operatorname{Ker} A$. □

We now give a description of all semidefinite H-polar decompositions (when they exist). The description of all possible H-nonnegative factors A is as follows.

THEOREM 5.5. *Let X be an $n \times n$ matrix that admits a semidefinite H-polar decomposition. Let S be an invertible matrix (over F) such that*

$$S^{-1}(X^{[*]}X)S = \operatorname{diag}(\lambda_i)_{i=1}^{\tau_1} \oplus 0_p \oplus \operatorname{diag}(\mu_i)_{i=1}^{\tau_2}, \tag{5.7}$$

$$S^*HS = -I_{\tau_1} \oplus H_0 \oplus I_{\tau_2}, \tag{5.8}$$

where $\lambda_i > 0$, $\mu_i > 0$, and

$$H_0 = \begin{bmatrix} 0 & 0 & I \\ 0 & H_2 & 0 \\ I & 0 & 0 \end{bmatrix}, \tag{5.9}$$

with respect to the decomposition

$$F^p = (\operatorname{Ker} X \cap (H \operatorname{Ker} X)^{\perp}) \oplus M_1 \oplus K, \tag{5.10}$$

where $\operatorname{Ker} X = (\operatorname{Ker} X \cap (H\operatorname{Ker} X)^{\perp}) \oplus M_1$. (*Such S exists by the proof of Theorem* 5.3.) *Then $X = UA$ for some H-unitary U and H-nonnegative A if and only if A has the form*

$$A = S\left(\operatorname{diag}\left(-\sqrt{\lambda_i}\right)_{i=1}^{\tau_1} \oplus A_0 \oplus \operatorname{diag}(\sqrt{\mu_i})_{i=1}^{\tau_2}\right) S^{-1},$$

where

$$A_0 = \begin{bmatrix} 0 & 0 & Y \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \in F^{p\times p}, \tag{5.11}$$

with respect to the decomposition (5.10), *and where Y is positive definite.*

Observe that by Theorem 5.3 the existence of S such that $S^{-1}(X^{[*]}X)S$ and S^*HS have the forms (5.7) and (5.8), respectively, is necessary for X to have a semidefinite H-polar decomposition.

Proof. By Proposition 5.2, $X = UA$ for some H-unitary U if and only if the H-nonnegative matrix A is such that $X^{[*]}X = A^2$ and $\operatorname{Ker} X = \operatorname{Ker} A$. These conditions are easily translated (using the invertibility of H_0 and H_2) into the statement of Theorem 5.5. ∎

For a fixed A, all possible H-unitary matrices U in the semidefinite H-polar decompositions $X = UA$ are given by an application of Theorem 2.4. This works as follows. Consider the decomposition of F^n,

$$F^n = N_1 \oplus N_2 \oplus N_3 \oplus N_4 \oplus N_5, \tag{5.12}$$

into five components as indicated in Theorem 5.5. With respect to this decomposition, let us write $S^{-1}US = [U_{ij}]_{i,j=1}^5$, $S^{-1}XS = [X_{ij}]_{i,j=1}^5$, $S^{-1}AS = A_1 \oplus A_0 \oplus A_5$. Assume that $X = UA$ and $X = \tilde{U}A$ are semidefinite H-polar decompositions of X. Also write $S^{-1}\tilde{U}S = [\tilde{U}_{ij}]_{i,j=1}^5$. Observing that A_1, A_5, and Y are invertible, we obtain from $X = UA = \tilde{U}A$ that

$$U_{j1} = \tilde{U}_{j1} = X_{j1}A_1^{-1}, \quad U_{j2} = \tilde{U}_{j2} = X_{j2}Y^{-1}, \quad U_{j5} = \tilde{U}_{j5} = X_{j5}A_5^{-1}.$$

Let $\hat{U} = \operatorname{col}[U_{j1}\ U_{j2}\ U_{j5}]_{j=1}^{5}$. Take $V_1 = N_1 \oplus N_2 \oplus N_5$ and $V_2 = UV = \tilde{U}V = \hat{U}V$. Then for all x, $y \in V_1$ we have

$$\langle H\hat{U}x, \hat{U}y\rangle = \langle Hx, y\rangle.$$

We see that both U and $\tilde{U}$ are Witt extensions of $\hat{U} : V_1 \to V_2$. Conversely, for any Witt extension V of $\hat{U}$ we have $X = VA$. Applying Theorem 2.4 to this situation gives the following.

THEOREM 5.6. *Suppose $X = UA$ is a semidefinite H-polar decomposition of X, and let A have the form as described in Theorem* 5.5, *with respect to the decomposition* (5.12) *of F^n. Then any H-unitary $\tilde{U}$ such that $X = \tilde{U}A$ is given by $\tilde{U} = UM$, where*

$$M = \begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & I & -P_2^* H_2 P_1 & P_3 - \frac{1}{2}P_2^* H_2 P_2 & 0 \\ 0 & 0 & P_1 & P_2 & 0 \\ 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix},$$

with respect to the decomposition (5.12). *Here P_2 is arbitrary, $P_3 = -P_3^*$, and P_1 is an arbitrary H_2-unitary matrix.*

In the real case, Theorem 2.6, together with Theorem 5.6, describes the number of connected components of $\mathcal{U}(H; \mathbf{R})$ from which the H-unitary factor in the semidefinite H-polar decompositions of X may be chosen.

COROLLARY 5.7 ($F = \mathbf{R}$). *Let X be an $n \times n$ matrix that admits a semidefinite H-polar decomposition. If $(H\ \operatorname{Ker} X)^{\perp} \supset \operatorname{Ker} X$, then all possible H-unitary factors in the semidefinite H-polar decompositions of X belong to the same connected component of $\mathcal{U}(U; \mathbf{R})$. Otherwise, let M_1 be the H-orthogonal complement of* $\operatorname{Ker} X \cap (H \operatorname{Ker} X)^{\perp}$ *in* $\operatorname{Ker} X$. *Then the H-unitary factors belong to two connected components of $\mathcal{U}(H; \mathbf{R})$ having determinants of opposite signs if $H|M_1$ is definite and to all four connected components of $\mathcal{U}(H; \mathbf{R})$ if $H|M_1$ is indefinite.*

The descriptions of the H-nonnegative and H-unitary factors in the polar decompositions of X obtained in Theorems 5.5 and 5.6, together with the real analytic structure of all Witt extensions (Theorems 2.5 and 2.6), allow one to describe the set of all possible H-polar decompositions of a given X in terms of a diffeomorphism (as a real analytic manifold). Using the results mentioned above, such a description is routine and is left to the interested readers.

6. Applications: Hyperbolic QR decompositions. The results of sections 2 and 4 have obvious applications to matrix equations of the form

$$A = UX, \tag{6.1}$$

where A is a given matrix, and X and U are matrices to be found such that U is H-unitary (usually additional requirements are imposed on X and/or U as well). Here A and X are $m \times n$ matrices over F (as usual, we assume that either $F = \mathbf{C}$ or $F = \mathbf{R}$), and H is an invertible $m \times m$ matrix over F which is either Hermitian or skew-symmetric (in the latter case we assume $F = \mathbf{R}$). Indeed, if U and V are solutions of (6.1) with the same A and X, then obviously $Ux = Vx$ for all x in the range of X. Thus, all H-unitary solutions of (6.1) can be treated as Witt extensions of $U|\operatorname{Range} X$, where U is one fixed H-unitary solution of (6.1). We will not explicitly present the straightforward statements that are obtained in this way. We focus instead on an important special case of equations (6.1) which is fundamental for a certain class

of algorithms for computing the eigenvalues of a matrix using the generalized Schur method, namely, hyperbolic QR decompositions (see, e.g., [B, OSB, V] and references therein).

In a typical version of hyperbolic QR decompositions, one seeks factorizations of the form (6.1), where $m \geq n$ and X is an upper triangular matrix

$$X = \begin{bmatrix} X_1 \\ 0 \end{bmatrix}$$

with invertible $n \times n$ matrix X_1. The factor U is H-unitary, where H is a fixed invertible Hermitian $m \times m$ matrix. Factorizations (6.1) of a given matrix A with the above properties will be called *hyperbolic QR decompositions* in this paper.

In what follows, we will use a basis $\{f_1, \dots, f_m\}$ in $\mathbf{F}^m$ such that $\{f_1, \dots, f_n\}$ forms a basis in $\begin{bmatrix} \mathbf{F}^n \\ 0 \end{bmatrix}$ and with respect to which the indefinite scalar product $[x, y] = \langle Hx, y\rangle$, $x, y \in \mathbf{F}^m$, induced by H has the Gramian matrix

$$\begin{bmatrix} 0 & 0 & I & 0 \\ 0 & J_1 & 0 & 0 \\ I & 0 & 0 & 0 \\ 0 & 0 & 0 & J_2 \end{bmatrix}, \tag{6.2}$$

where I is the $n_0 \times n_0$ identity matrix and J_1 is the diagonal $(n_+ + n_-) \times (n_+ + n_-)$ matrix having the first n_+ diagonal elements equal to $+1$ and the remaining n_- diagonal elements equal to -1; here $n_0 + n_+ + n_- = n$. Similarly, J_2 is a diagonal matrix with entries $+1$ and -1 on the main diagonal. (Compare with (2.5).) A basis $\{f_1, \dots, f_m\}$ with the above properties will be called *admissible*.

THEOREM 6.1. *Let $A = U_0 X_0$ be a hyperbolic QR decomposition of a given $m \times n$ matrix A. Then every hyperbolic QR decomposition $A = \tilde{U} X_0$ of A with the same factor X_0 is given by the following formula, written as a block 4×4 matrix (compatible with (6.2)) with respect to an admissible basis: $\tilde{U} = U_0 M$, where*

$$M = \begin{bmatrix} I & 0 & -\frac{1}{2} P_2^* J_2 P_2 + P_3 & -P_2^* J_2 P_1 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & P_2 & P_1 \end{bmatrix}.$$

Here P_1 is J_2-unitary, P_3 is a skew-self-adjoint $n_0 \times n_0$ matrix, and P_2 is an arbitrary $(m - n - n_0) \times n_0$ matrix.

The proof is a straightforward application of Theorem 2.4.

Applying Theorem 2.6, we have the following corollary in the real case.

COROLLARY 6.2. *The set of all hyperbolic QR decompositions $A = \tilde{U} X_0$ of a given $m \times n$ matrix A with a given $m \times n$ factor X_0 is connected if*

$$\pi(H) = n_+ + n_0, \qquad \nu(H) = n_- + n_0$$

has two connected components if exactly one of the numbers $p = \pi(H) - n_+ - n_0$ and $q = \nu(H) - n_- - n_0$ is positive and has four connected components if both p and q are positive.

We do not discuss here the problem of existence of hyperbolic QR decompositions for a given $m \times n$ matrix A and a given invertible Hermitian $m \times m$ matrix H and only mention that the obvious necessary condition for A to have full column rank is

not sufficient. A characterization of all square matrices A that admit a decomposition $A = UX$, where U is H_1-H_2-unitary and X is upper triangular and nonsingular, is given in Theorem 2.3 of [B]. (The paper [B] considers only diagonal matrices H_1 and H_2, which is the most important case for the development of algorithms based on the generalized Schur method.) An extension to the case of rectangular matrices is presented in [V].

REFERENCES

[A] E. Artin, *Geometric Algebra.* Interscience Publishers, New York, 1957.

[AI1] T. Ya. Azizov and I. S. Iohvidov, *Linear Operators in Spaces with an Indefinite Metric*, John Wiley and Sons, New York, 1989.

[AI2] T. Ya. Azizov and E. I. Iohvidov, *The development of some of V. P. Potapov's ideas in the geometric theory of operators in spaces with indefinite metric*, in Matrix and Operator Valued Functions, Operator Theory, Vol. 72, I. Gohberg and L.A. Sakhnovich, eds., Birkhäuser, Basel, 1994, pp. 17–27.

[Bo] J. Bognár, *Indefinite Inner Product Spaces,* Springer, Berlin, 1974.

[BMRRR1] Y. Bolshakov, C. V. M. van der Mee, A. C. M. Ran, B. Reichstein, and L. Rodman, *Polar decompositions in finite dimensional indefinite scalar product spaces: General theory*, Linear Algebra Appl., to appear.

[BMRRR2] Y. Bolshakov, C. V. M. van der Mee, A. C. M. Ran, B. Reichstein, and L. Rodman, *Polar decompositions in finite dimensional indefinite scalar product spaces: Special cases and applications*, in Recent Developments in Operator Theory and its Applications, Operator Theory, Vol. 87, I. Gohberg, P. Lancaster, and P. N. Shivakumar, eds., Birkhäuser, Basel, 1996, pp. 61–94. Errata, Integral Equations and Operator Theory, to appear.

[BR] Y. Bolshakov and B. Reichstein, *Unitary equivalence in an indefinite scalar product: An analogue of singular value decomposition*, Linear Algebra Appl., 222 (1995), pp. 155–226.

[B] A. Bunse-Gerstner, *An analysis of the HR algorithm for computing the eigenvalues of a matrix*, Linear Algebra Appl., 35 (1981), pp. 155–173.

[C] C. Chevalley, *Theory of Lie Groups* I, Princeton University Press, Princeton, NJ, 1946.

[CH] D. Choudhury and R. A. Horn, *A complex orthogonal-symmetric analog of the polar decomposition*, SIAM J. Alg. Disc. Meth., 8 (1987), pp. 219–225.

[D] J. Dieudonné, *La Géométrie des Groupes Classiques*, Springer, Berlin, 1955.

[GLR] I. Gohberg, P. Lancaster, and L. Rodman, *Matrices and Indefinite Scalar Products*, Operator Theory, Vol. 8, Birkhäuser, Basel, 1983.

[HM1] R. A. Horn and D. I. Merino, *Contragredient equivalence: A canonical form and some applications*, Linear Algebra Appl., 214 (1995), pp. 43–92.

[HM2] R. A. Horn and D. I. Merino, *A real coninvolutory analog of the polar decomposition*, Linear Algebra Appl., 190 (1993), pp. 209–227.

[IKL] I. S. Iohvidov, M. G. Krein, and H. Langer, *Introduction to the Spectral Theory of Operators in Spaces with an Indefinite Metric*, Akademie-Verlag, Berlin, 1982.

[K] I. Kaplansky, *Algebraic polar decomposition*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 213–217.

[KS1] M. G. Krein and Ju. L. Shmul'jan, *J-polar representation of plus operators*, Mat. Issled., 1 (1966), pp. 172–210. (In Russian.) English translation: Amer. Math. Soc. Transl., Ser. 2, 85 (1969), pp. 115–143.

[KS2] M. G. Krein and Ju. L. Shmul'jan, *On plus operators in a space with an indefinite metric*, Mat. Issled., 1 (1966), pp. 131–161. (In Russian.) English translation: Amer. Math. Soc. Transl., Ser. 2, 85 (1969), pp. 93–113.

[M] C. V. M. van der Mee, *An eigenvalue criterion for matrices transforming Stokes parameters*, J. Math. Phys., 34 (1993), pp. 5072–5088.

[MH] C. V. M. van der Mee and J. W. Hovenier, *Structure of matrices transforming Stokes parameters*, J. Math. Phys., 33 (1992), pp. 3574–3584.

[OSB] R. Onn, A. O. Steinhardt, and A. W. Bojanczyk, *The hyperbolic singular value decomposition and applications*, IEEE Trans. Signal Proc., 39 (1991), pp. 1575–1588.

[P1] V. P. Potapov, *Multiplicative structure of J-nonexpansive matrix functions*, Trudy Moskov Mat. Obshch., 4 (1955), pp. 125–236. (In Russian.) English translation: Amer. Math. Soc. Transl., Ser. 2, 15 (1960), pp. 131–243.

[P2] V. P. Potapov, *A theorem on the modulus*, I. *Main concepts. The modulus*, *Theory of Functions and Functional Analysis*, 38 (1982), pp. 91–101, 129. (In Russian.) English Translation: Amer. Math. Soc. Transl., Ser. 2, Vol. 138, pp. 55–65.

[T] R. C. Thompson, *Pencils of complex and real symmetric and skew matrices*, Linear Algebra Appl., 147 (1991), pp. 323–371.

[V] A.-J. van der Veen, *A Schur method for low-rank matrix approximation*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 139–160.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 3, pp. 775–791, July 1997

013

PERTURBATION ANALYSES FOR THE QR FACTORIZATION*

XIAO-WEN CHANG†, CHRISTOPHER C. PAIGE†, AND G. W. STEWART‡

Abstract. This paper gives perturbation analyses for Q_1 and R in the QR factorization $A = Q_1R$, $Q_1^TQ_1 = I$ for a given real $m \times n$ matrix A of rank n and general perturbations in A which are sufficiently small in norm. The analyses more accurately reflect the sensitivity of the problem than previous such results. The condition numbers here are altered by any column pivoting used in $AP = Q_1R$, and the condition number for R is bounded for a fixed n when the standard column pivoting strategy is used. This strategy also tends to improve the condition of Q_1, so the computed Q_1 and R will probably both have greatest accuracy when we use the standard column pivoting strategy.

First-order perturbation analyses are given for both Q_1 and R. It is seen that the analysis for R may be approached in two ways—a detailed "matrix–vector equation" analysis which provides a tight bound and corresponding condition number, which unfortunately is costly to compute and not very intuitive, and a simpler "matrix equation" analysis which provides results that are usually weaker but easier to interpret and which allows the efficient computation of satisfactory estimates for the actual condition number. These approaches are powerful general tools and appear to be applicable to the perturbation analysis of any matrix factorization.

Key words. QR factorization, perturbation analysis, condition estimation, matrix equations, pivoting

AMS subject classifications. 15A23, 65F35

PII. S0895479896297720

1. Introduction.

The QR factorization is an important tool in matrix computations (see, for example, [5]): given an $m \times n$ real matrix A with full column rank, there exists a unique $m \times n$ real matrix Q_1 with orthonormal columns and a unique nonsingular upper triangular $n \times n$ real matrix R with positive diagonal entries so

$$A = Q_1R.$$

The matrix Q_1 is referred to as the orthogonal factor and R the triangular factor.

Whenever Q_1 or R has meaning in its own right, we will be interested in how sensitive it is to changes in the original matrix A. Some practical examples where this sensitivity is important are described in [3]. Here we give perturbation analyses leading to condition numbers. Since a condition number (as a function of a matrix of a certain class) must be from a bound which is attainable (for any matrix in the given class) we will use this rigorous terminology and use qualified terms (e.g., "condition estimate," "condition bound") when this criterion is not met.

Suppose $A(t) \equiv A + tG$ has the unique QR factorization $A(t) = Q_1(t)R(t)$. If we differentiate $R(t)^TR(t) = A(t)^TA(t)$ with respect to t and set $t = 0$ we have

$$R^T\dot{R}(0) + \dot{R}^T(0)R = A^T\dot{A}(0) + \dot{A}(0)^TA = R^TQ_1^TG + G^TQ_1R, \tag{1.1}$$

* Received by the editors January 22, 1996; accepted for publication (in revised form) by N. J. Higham August 23, 1996.

http://www.siam.org/journals/simax/18-3/29772.html

† School of Computer Science, McGill University, Montreal, Quebec, Canada H3A 2A7 (chang@cs.mcgill.ca, chris@cs.mcgill.ca). The research of these authors was supported by NSERC of Canada grant OGP0009236.

‡ Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 (stewart@cs.umd.edu). The research of this author was supported in part by National Science Foundation grant CCR 95503126.

which with given A and G is a linear equation for the upper triangular matrix $\dot{R}(0)$. But $\dot{R}(0)$ determines the sensitivity of $R(t)$ at $t = 0$, and so the core of any first-order perturbation analysis for the QR factorization is the use of (1.1) to determine, or to bound, $\dot{R}(0)$. We first discuss two main ways of approaching this problem.

Chang [1] pointed out that most of the published results on the sensitivity of factorizations, such as LU, Cholesky, and QR, were extremely weak for certain classes of matrices and showed why the condition was often significantly improved by pivoting. To do this he created a general approach for obtaining provably tight results and corresponding condition numbers for such problems. We will call this the "matrix–vector equation" approach. For the QR factorization this involves expressing (1.1) as a matrix–vector equation of the form

$$W_R \mathrm{uvec}(\dot{R}(0)) = Z_R \mathrm{vec}(Q_1^T G),$$

where W_R and Z_R are matrices involving the elements of R and $\mathrm{vec}(\cdot)$ transforms its argument into a vector of its elements. The notation $\mathrm{uvec}(\cdot)$ denotes a variant of $\mathrm{vec}(\cdot)$ defined in section 5.2. Previously, the most used approach to perturbation analyses of factorizations was what we will call the "matrix equation" approach, which keeps equations like (1.1) in their matrix–matrix form. Stewart [13] used a construct, partly illustrated by the "up" and "low" notation in section 2, which makes the matrix equation approach a more usable and intuitive tool. He combined this with scaling to produce new matrix equation analyses which are straightforward and provide greater insight into the sensitivities of the problems. These new matrix equation analyses do not in general provide tight results like the matrix–vector equation analyses do, but they are usually more simple and provide practical estimates for the condition numbers obtained from the latter. A combination of the two approaches provides a full understanding of the cases we have examined so far and is a powerful general tool that appears to be applicable to the perturbation analysis of any matrix factorization.

The perturbation analysis for the QR factorization has been considered by several authors. The first norm-based result for R was presented by Stewart [11]. That was further modified and improved by Sun [15]. Using different approaches Sun [15] and Stewart [12] gave *first-order* norm-based perturbation analyses for R. A first-order so-called "componentwise" perturbation analysis for R was given by Zha [19] (this assumes ΔA has the form of the equivalent backward rounding error from a numerically stable computation of the QR factorization), and a *strict* analysis for the components of R was given by Sun [16]. These papers also gave analyses for Q_1. More recently Sun [17] gave strict perturbation bounds for Q_1 alone.

The purpose of this paper is to establish new first-order perturbation bounds which are generally sharper than the equivalent results for the R factor in [12, 15], and more straightforward than the sharp result in [17] for the Q_1 factor.

In section 2 we define some notation and give a result we will use throughout the paper. In section 3 we survey important key results on the sensitivity of R and Q_1 which will be useful later. In section 4 we give a refined perturbation analysis for Q_1, showing in a simple way why the standard column pivoting strategy for A can be beneficial for certain aspects of the sensitivity of Q_1. In section 5 we analyze the perturbation in R, first by the straightforward matrix equation approach, then by the more detailed and tighter matrix–vector equation approach. We give numerical results and suggest practical condition estimators in section 6, and summarize and comment on our findings in section 7.

2. Notation and basics. To simplify the presentation, for any $n \times n$ matrix X, we define the upper and lower triangular matrices

$$\text{up}(X) \equiv \begin{pmatrix} \frac{1}{2}x_{11} & x_{12} & \cdot & x_{1n} \\ 0 & \frac{1}{2}x_{22} & \cdot & x_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \frac{1}{2}x_{nn} \end{pmatrix}, \qquad \text{low}(X) \equiv \text{up}(X^T)^T, \tag{2.1}$$

so that $X = \text{low}(X) + \text{up}(X)$. For general X

$$\|\text{low}(X) - [\text{low}(X)]^T\|_F, \leq \sqrt{2}\|X\|_F. \tag{2.2}$$

For symmetric X

$$2\|\text{low}(X)\|_F^2 = 2\|\text{up}(X)\|_F^2 = \|X\|_F^2 - \frac{1}{2}(x_{11}^2 + x_{22}^2 + \cdots + x_{nn}^2) \leq \|X\|_F^2. \tag{2.3}$$

To illustrate a basic use of "up," we show that for any given $n \times n$ nonsingular upper triangular R and any given $n \times n$ symmetric M, the equation of the form (cf. (1.1))

$$R^T U + U^T R = M \tag{2.4}$$

always has a unique upper triangular solution U. Since UR^{-1} is upper triangular in $UR^{-1} + (UR^{-1})^T = R^{-T}MR^{-1}$ symmetric, we see immediately that $UR^{-1} = \text{up}(R^{-T}MR^{-1})$, so UR^{-1} and therefore U is uniquely defined. We will describe other uses later.

Our perturbation bounds for Q_1 will be tighter if we bound separately the perturbations along the column space of A and along its orthogonal complement. Thus we introduce the following notation. For real $m \times n$ A, let P_1 be the orthogonal projector onto $\mathcal{R}(A)$ and P_2 be the orthogonal projector onto $\mathcal{R}(A)^\perp$. For real $m \times n$ ΔA define

$$\epsilon \equiv \|\Delta A\|_F/\|A\|_2, \quad \epsilon_1 \equiv \|P_1\Delta A\|_F/\|A\|_2, \quad \epsilon_2 \equiv \|P_2\Delta A\|_F/\|A\|_2, \tag{2.5}$$

so $\epsilon^2 = \epsilon_1^2 + \epsilon_2^2$. When $\epsilon > 0$ in (2.5) we also define

$$G \equiv \Delta A/\epsilon, \tag{2.6}$$

so for the QR factorization $A = Q_1 R$

$$\|G\|_F = \|A\|_2 = \|R\|_2. \tag{2.7}$$

We will use the following standard result.

LEMMA 2.1. *For real $m \times n$ A with rank n, real $A + \Delta A$ has rank n if*

$$\kappa_2(A)\frac{\|P_1\Delta A\|_2}{\|A\|_2} < 1, \tag{2.8}$$

where $\kappa_2(A) \equiv \|A^\dagger\|_2\|A\|_2$ and P_1 is the orthogonal projector onto $\mathcal{R}(A)$.

Proof. Let A have the QR factorization

$$A = Q\begin{pmatrix} R \\ 0 \end{pmatrix} = (Q_1, Q_2)\begin{pmatrix} R \\ 0 \end{pmatrix}; \quad \text{then} \quad Q^T(A + \Delta A) = \begin{pmatrix} R + Q_1^T\Delta A \\ Q_2^T\Delta A \end{pmatrix},$$

which necessarily has full column rank if $\|Q_1^T\Delta A\|_2 < \sigma_{\min}(A)$, the smallest singular value of A. But this inequality is just (2.8). □

COROLLARY 2.2. *If nonzero ΔA satisfies* (2.8)*, then for ϵ and G defined in* (2.5) *and* (2.6) *$A + tG$ has full column rank and therefore a unique QR factorization for all $|t| \leq \epsilon$.* □

3. Previous norm-based results. In this section we summarize the strongest norm-based results by previous authors. We first give a derivation of what is essentially Sun's [15] and Stewart's [12] first-order norm-based perturbation result for R, since the techniques, intermediate equations, and results will be useful later.

THEOREM 3.1 (see [15]). *Let $A \in \mathcal{R}^{m\times n}$ have full column rank and QR factorization $A = Q_1 R$, and let ΔA be a real $m \times n$ matrix. Define $\epsilon \equiv \|\Delta A\|_F/\|A\|_2$ and $\epsilon_1 \equiv \|Q_1^T \Delta A\|_F/\|A\|_2$; see* (2.5). *If* (2.8) *holds then $A + \Delta A$ has a unique QR factorization*

$$A + \Delta A = (Q_1 + \Delta Q_1)(R + \Delta R), \tag{3.1}$$

where

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \sqrt{2}\kappa_2(A)\epsilon_1 + O(\epsilon^2). \tag{3.2}$$

Proof. Let $G \equiv \Delta A/\epsilon$ (if $\epsilon = 0$ the theorem is trivial). From Corollary 2.2 $A+tG$ has the unique QR factorization

$$A(t) \equiv A + tG = Q_1(t)R(t) \quad \text{for all } |t| \le \epsilon, \quad \text{where} \tag{3.3}$$

$$Q_1^T(t)Q_1(t) = I. \tag{3.4}$$

Notice that $R(0) = R$ and $R(\epsilon) = R + \Delta R$.

It is easy to verify that $Q_1(t)$ and $R(t)$ are twice continuously differentiable for $|t| \le \epsilon$ from the algorithm for the QR factorization. Thus as in (1.1) we have

$$R^T\dot{R}(0) + \dot{R}^T(0)R = R^T Q_1^T G + G^T Q_1 R, \tag{3.5}$$

which (see (2.4)) is a linear equation *uniquely* defining the elements of $\dot{R}(0)$ in terms of the elements of $Q_1^T G$. From upper triangular $\dot{R}(0)R^{-1}$ in

$$\dot{R}(0)R^{-1} + (\dot{R}(0)R^{-1})^T = Q_1^T G R^{-1} + (Q_1^T G R^{-1})^T, \tag{3.6}$$

we see with (2.1) that

$$\dot{R}(0) = \mathrm{up}[Q_1^T G R^{-1} + (Q_1^T G R^{-1})^T]R, \tag{3.7}$$

so with (2.3)

$$\begin{aligned}\|\dot{R}(0)\|_F &\le \frac{1}{\sqrt{2}}\|Q_1^T G R^{-1} + (Q_1^T G R^{-1})^T\|_F \|R\|_2 \\ &\le \sqrt{2}\|Q_1^T G R^{-1}\|_F \|R\|_2 \le \sqrt{2}\kappa_2(R)\|Q_1^T G\|_F,\end{aligned}$$

and since from (2.5)–(2.7) $\|Q_1^T G\|_F = \|A\|_2\epsilon_1/\epsilon$, and $\|R^{-1}\|_2 = \|A^\dagger\|_2$,

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_2} \le \sqrt{2}\kappa_2(A)\epsilon_1/\epsilon. \tag{3.8}$$

The Taylor expansion for $R(t)$ about $t = 0$ gives at $t = \epsilon$

$$R + \Delta R = R(\epsilon) = R(0) + \epsilon\dot{R}(0) + O(\epsilon^2), \tag{3.9}$$

so that

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \frac{\|\dot{R}(0)\|_F}{\|R\|_2}\epsilon + O(\epsilon^2),$$

which, combined with (3.8), gives (3.2). □

This proof shows that the key point in deriving a first-order perturbation bound for R is the use of (3.5) to give a good bound on the sensitivity $\|\dot{R}(0)\|_F/\|R\|_2$. Since we obtained the bounds directly from (3.7), this was a "matrix equation" analysis.

We now show how a recent perturbation result for Q_1 given by Sun [17] can be obtained in the present setting, since the analysis can easily be extended to obtain a more refined result in a simple way. Note that the hypotheses of the following theorem are those of Theorem 3.1, so we can use results from the latter theorem.

THEOREM 3.2 (see [17]). *Let $A \in \mathcal{R}^{m\times n}$ be of full column rank, with the QR factorization $A = Q_1R$, and let ΔA be a real $m \times n$ matrix. If $\epsilon \equiv \|\Delta A\|_F/\|A\|_2$ and* (2.8) *holds, then $A + \Delta A$ has a unique QR factorization*

$$A + \Delta A = (Q_1 + \Delta Q_1)(R + \Delta R),$$

where

$$\|\Delta Q_1\|_F \le \sqrt{2}\kappa_2(A)\epsilon + O(\epsilon^2). \tag{3.10}$$

Proof. Let $G \equiv \Delta A/\epsilon$ (if $\epsilon = 0$ the theorem is trivial). From Corollary 2.2 $A+tG$ has the unique QR factorization $A(t) \equiv A + tG = Q_1(t)R(t)$ with $Q_1(t)^TQ_1(t) = I$ for all $|t| \le \epsilon$. Differentiating these at $t = 0$ gives

$$G = Q_1\dot{R}(0) + \dot{Q}_1(0)R, \qquad Q_1^T\dot{Q}_1(0) \quad \text{skew symmetric.}$$

It follows that

$$\dot{Q}_1(0) = GR^{-1} - Q_1\dot{R}(0)R^{-1},$$

so with any Q_2 such that $Q \equiv (Q_1, Q_2)$ is square and orthogonal,

$$\begin{aligned} Q_2^T\dot{Q}_1(0) &= Q_2^TGR^{-1}, \\ Q_1^T\dot{Q}_1(0) &= Q_1^TGR^{-1} - \dot{R}(0)R^{-1}. \end{aligned} \tag{3.11}$$

Now using (2.1), we have with (3.7) in the proof of Theorem 3.1 that

$$\begin{aligned} Q_1^T\dot{Q}_1(0) &= \mathrm{low}(Q_1^TGR^{-1}) + \mathrm{up}(Q_1^TGR^{-1}) \\ &\quad -\mathrm{up}[Q_1^TGR^{-1} + (Q_1^TGR^{-1})^T] \\ &= \mathrm{low}(Q_1^TGR^{-1}) - [\mathrm{low}(Q_1^TGR^{-1})]^T. \end{aligned} \tag{3.12}$$

We see from this, (2.2), (3.11), and $\|G\|_F = \|A\|_2$ from (2.7) that

$$\begin{aligned} \|\dot{Q}_1(0)\|_F^2 &= \|Q_1^T\dot{Q}_1(0)\|_F^2 + \|Q_2^T\dot{Q}_1(0)\|_F^2 \\ &\le 2\|Q_1^TGR^{-1}\|_F^2 + \|Q_2^TGR^{-1}\|_F^2 \le 2\|GR^{-1}\|_F^2, \\ \|\dot{Q}_1(0)\|_F &\le \sqrt{2}\|R^{-1}\|_2\|G\|_F = \sqrt{2}\kappa_2(A), \end{aligned} \tag{3.13}$$

and from the Taylor expansion for $Q_1(t)$ about $t = 0$ at $t = \epsilon$,

$$Q_1 + \Delta Q_1 = Q_1(\epsilon) = Q_1(0) + \epsilon\dot{Q}_1(0) + O(\epsilon^2), \tag{3.14}$$

so that $\|\Delta Q_1\|_F \le \epsilon\|\dot{Q}_1(0)\|_F + O(\epsilon^2)$, which with (3.13) gives (3.10). □

4. Refined analysis for Q_1. The results of Sun [17] give about as good as possible overall bounds on the change ΔQ_1 in Q_1. But by looking at how ΔQ_1 is distributed between $\mathcal{R}(Q_1)$ and its orthogonal complement, and following the ideas in Theorem 3.2, we are able to obtain a result which is tight but, unlike the related tight result in [17], easy to follow. It makes clear exactly where any ill conditioning lies. From (3.14) with $Q = (Q_1, Q_2)$ square and orthogonal,

$$\Delta Q_1 = \epsilon Q_1 Q_1^T \dot{Q}_1(0) + \epsilon Q_2 Q_2^T \dot{Q}_1(0) + O(\epsilon^2),$$

and the key is to bound the first term on the right separately from the second. Note from (3.11) with (2.5)–(2.7) that

$$\|Q_2^T \dot{Q}_1(0)\|_F = \|Q_2^T G R^{-1}\|_F \le \|R^{-1}\|_2 \|Q_2^T G\|_F = \kappa_2(A)\epsilon_2/\epsilon,$$

where G can be chosen to give equality here for any given A. Hence $\kappa_2(A)$ is the condition number for that part of ΔQ_1 in $\mathcal{R}(Q_2)$:

$$\|Q_2^T \Delta Q_1\|_F \le \kappa_2(A)\epsilon_2 + O(\epsilon^2). \tag{4.1}$$

Now we turn to the part of ΔQ_1 in $\mathcal{R}(Q_1)$. We see from (3.12) that $n \times n$

$$S \equiv Q_1^T \dot{Q}_1(0) = \mathrm{low}(Q_1^T G R^{-1}) - [\mathrm{low}(Q_1^T G R^{-1})]^T, \tag{4.2}$$

which is skew symmetric with clearly zero diagonal. Thus if $n = 1$, $S = Q_1^T \dot{Q}_1(0) = 0$. For $n > 1$ let R_j and S_j denote the leading $j \times j$ blocks of R and S, respectively, G_j the matrix of the first j columns of G and $Q_1 = [q_1, \ldots, q_n]$. If we write

$$S_1 = 0, \qquad S_j = \left(\begin{array}{c|c} S_{j-1} & -s_j \\ \hline s_j^T & 0 \end{array}\right), \quad j = 2, \ldots, n,$$

where s_j has j–1 elements, then from the upper triangular form of R in (4.2)

$$\begin{aligned} s_j^T &= q_j^T G_{j-1} R_{j-1}^{-1}, \\ \frac{1}{2}\|S\|_F^2 &= \|s_2\|_2^2 + \cdots + \|s_n\|_2^2 \\ &\le \|R_1^{-1}\|_2^2 \|G_1^T q_2\|_2^2 + \cdots + \|R_{n-1}^{-1}\|_2^2 \|G_{n-1}^T q_n\|_2^2 \\ &\le \|R_{n-1}^{-1}\|_2^2 \|G_{n-1}^T Q_1\|_F^2 \le \|R_{n-1}^{-1}\|_2^2 \|Q_1^T G\|_F^2. \end{aligned}$$

Clearly for any R_{n-1} equality is obtained by taking $G = (q_n y^T, 0)$, with y nonzero such that $\|R_{n-1}^{-T} y\|_2 = \|R_{n-1}^{-1}\|_2 \|y\|_2$. It follows that the bound is tight in

$$\begin{aligned} \|Q_1^T \dot{Q}_1(0)\|_F = \|S\|_F &\le \sqrt{2}\|R_{n-1}^{-1}\|_2 \|A\|_2 \epsilon_1/\epsilon, \\ \|Q_1^T \Delta Q_1\|_F &\le \sqrt{2}\|R_{n-1}^{-1}\|_2 \|A\|_2 \epsilon_1 + O(\epsilon^2), \end{aligned} \tag{4.3}$$

so the condition number for that part of ΔQ_1 in $\mathcal{R}(Q_1)$ is not $\sqrt{2}\kappa_2(A)$ but

$$\kappa_{Q_1}(A) \equiv \sqrt{2}\|R_{n-1}^{-1}\|_2 \|A\|_2. \tag{4.4}$$

In some problems we are mainly (in fact *only*, if A is square and nonsingular) interested in the change in Q_1 lying in $\mathcal{R}(Q_1)$, and this result shows its bound can be smaller than we previously thought. In particular, if A has only one small singular value and we use the standard column pivoting strategy in computing the QR factorization, then R_{n-1} will usually be quite well conditioned compared with R and we will have $\|R_{n-1}^{-1}\|_2 \|A\|_2 \ll \kappa_2(A)$. However, for some special cases this may not be true (for example, the Kahan matrix in section 6), and then a rank revealing pivoting strategy such as in [8] would be required to obtain such an improvement.

5. Perturbation analyses for *R*. In section 3 we saw that the key to deriving first-order perturbation bounds for R in the QR factorization of full column rank A is equation (3.5), which has the general form of finding (bounding) X in terms of given R and F in the matrix equation

$$R^T X + X^T R = R^T F + F^T R, \quad X \text{ and } R \text{ upper triangular, } R \text{ nonsingular.} \tag{5.1}$$

Sun [15] and Stewart [12] originally analyzed this using the matrix equation approach to give the result in Theorem 3.1. We will now analyze it in two new ways. The first, the refined matrix equation approach, gives a clear improvement to Theorem 3.1, while the second, the matrix–vector equation approach, gives a further improvement still—provably tight bounds leading to the condition number $\kappa_R(A)$ for R in the QR factorization of A. Both approaches provide efficient condition estimators (see [2] for the matrix–vector equation approach) and nice results for the special case of $AP = Q_1 R$, where P is a permutation matrix giving the standard column pivoting, but we will derive only the matrix equation versions of these. The tighter but more complicated matrix–vector equation analysis for the case of pivoting is given in [2], and only the results will be quoted here.

5.1. Refined matrix equation analysis for *R*. Our proof of Theorem 3.1 used (3.5) to produce the matrix equation (3.7) and derived the bounds directly from this. We now look at this approach more closely, but at first using the general form (5.1) to keep our thinking clear. From this we see that

$$X = \mathrm{up}(FR^{-1} + R^{-T}F^T)R.$$

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 $R = D\bar{R}$. Note that for any matrix B we have $\mathrm{up}(B)D = \mathrm{up}(BD)$. Hence if we define $B \equiv F\bar{R}^{-1}$, then

$$X = \mathrm{up}(F\bar{R}^{-1} + D^{-1}\bar{R}^{-T}F^T D)\bar{R} = [\mathrm{up}(B) + D^{-1}\mathrm{up}(B^T)D]\bar{R}. \tag{5.2}$$

With obvious notation, the upper triangular matrix $\mathrm{up}(B) + D^{-1}\mathrm{up}(B^T)D$ has (i,j) element $\beta_{ij} + \beta_{ji}\delta_j/\delta_i$ for $i < j$ and (i,i) element β_{ii}. To bound this, we use the following lemma.

LEMMA 5.1. *For $n \times n$ B and $D \in \mathcal{D}_n$,*

$$\phi \equiv \|\mathrm{up}(B) + D^{-1}\mathrm{up}(B^T)D\|_F \le \sqrt{1 + \zeta_D^2}\,\|B\|_F, \tag{5.3}$$

where

$$\zeta_D \equiv \max_{1 \le i < j \le n} \{\delta_j/\delta_i\}. \tag{5.4}$$

Proof. Clearly,

$$\phi^2 = \sum_{i=1}^{n} \beta_{ii}^2 + \sum_{j=2}^{n}\sum_{i=1}^{j-1} \left(\beta_{ij} + \frac{\delta_j}{\delta_i}\beta_{ji}\right)^2.$$

But by the Cauchy–Schwarz theorem, $(\beta_{ij} + \frac{\delta_j}{\delta_i}\beta_{ji})^2 \le (\beta_{ij}^2 + \beta_{ji}^2)(1 + (\frac{\delta_j}{\delta_i})^2)$, so

$$\phi^2 \le \sum_{i=1}^{n} \beta_{ii}^2 + \sum_{j=2}^{n}\sum_{i=1}^{j-1} (\beta_{ij}^2 + \beta_{ji}^2)\left(1 + \left(\frac{\delta_j}{\delta_i}\right)^2\right)$$

$$= \|B\|_F^2 + \sum_{j=2}^{n}\sum_{i=1}^{j-1}(\beta_{ij}^2 + \beta_{ji}^2)\left(\frac{\delta_j}{\delta_i}\right)^2$$

$$\leq \|B\|_F^2 + \zeta_D^2\|B\|_F^2. \qquad \square \tag{5.5}$$

We can now bound the solution X of (5.1):

$$\|X\|_F \leq \phi \cdot \|\bar{R}\|_2 \leq \sqrt{1+\zeta_D^2}\|B\|_F\|\bar{R}\|_2 = \sqrt{1+\zeta_D^2}\|F\bar{R}^{-1}\|_F\|\bar{R}\|_2 \tag{5.6}$$

$$\leq \sqrt{1+\zeta_D^2}\kappa_2(\bar{R})\|F\|_F. \tag{5.7}$$

But this is true for all $D \in \mathcal{D}_n$, so for the upper triangular solution X of (5.1)

$$\|X\|_F \leq \kappa_{ME}(A)\|F\|_F, \tag{5.8}$$

$$\kappa_{ME}(A) \equiv \inf_{D\in\mathcal{D}_n} \kappa(R, D), \tag{5.9}$$

$$\kappa(R, D) \equiv \sqrt{1+\zeta_D^2}\kappa_2(D^{-1}R), \tag{5.10}$$

where ζ_D is defined in (5.4). This gives the encouraging result

$$\kappa_{ME}(A) \leq \kappa(R, I) = \sqrt{2}\kappa_2(R) = \sqrt{2}\kappa_2(A). \tag{5.11}$$

Comparing (3.5) with (5.1), we see for the QR factorization, with (2.5)–(2.7),

$$\|\dot{R}(0)\|_F \leq \kappa_{ME}(A)\|Q_1^T G\|_F = \kappa_{ME}(A)\|A\|_2\epsilon_1/\epsilon.$$

Hence

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_2} \leq \kappa_{ME}(A)\epsilon_1/\epsilon, \tag{5.12}$$

and from (3.9) for a change $\Delta A = \epsilon G$ in A we have

$$\frac{\|\Delta R\|_F}{\|R\|_2} \leq \kappa_{ME}(A)\epsilon_1 + O(\epsilon^2), \tag{5.13}$$

where from (5.11) these are never worse than the bound in Theorem 3.1.

With the standard column pivoting strategy in $AP = Q_1 R$, P a permutation matrix, this analysis leads to a very nice result. Here the elements of R satisfy

$$r_{ii}^2 \geq \sum_{k=i}^{j} r_{kj}^2, \qquad i = 1, \ldots, n, \quad j = i, \ldots, n,$$

so $r_{11}^2 \geq r_{22}^2 \geq \cdots \geq r_{nn}^2$. If D is the diagonal of R then $\zeta_D \leq 1$, and from (5.9) and (5.10)

$$\kappa_{ME}(AP) \leq \kappa(R, D) \leq \sqrt{2}\kappa_2(\bar{R}), \qquad \bar{R} = D^{-1}R.$$

But then $1 = |\bar{r}_{ii}| \geq |\bar{r}_{ij}|$ for all $j \geq i$, and it follows from [7, Theorem 8.13] that

$$1 \leq \|\bar{R}^{-1}\|_2 \leq 2^{n-1},$$

so since

$$\|\bar{R}\|_2^2 \leq \|\bar{R}\|_F^2 \leq n(n+1)/2,$$

$$\kappa_2(\bar{R}) \leq 2^{n-1}\cdot\sqrt{n(n+1)/2},$$

$$\kappa_{ME}(AP) \leq \sqrt{2}\kappa_2(\bar{R}) \leq 2^{n-1}\sqrt{n(n+1)}. \tag{5.14}$$

So with the standard pivoting strategy, the sensitivity of R is bounded for any n.

Remark 5.1. Clearly $\kappa_{ME}(A)$ is a potential candidate for the condition number of R in the QR factorization. From (5.9), $\kappa_{ME}(A)$ depends solely on R, but it will only be the condition number if for *any* nonsingular upper triangular R we can find an F in (5.1) giving equality in (5.8). From (5.7) this can only be true if every column of F^T lies in the space of the right singular vectors corresponding to the maximum singular value of $\bar{R}^{-T}$. Such a restriction is in general too strong for (5.6) to be made an equality as well (see the lead up to (5.5)). But for a class of R this *is* possible. If R is diagonal, we can take $D = R$ giving $\bar{R} = I$, and the first restriction on F disappears. Let i and j be such that $\zeta_D = \delta_j/\delta_i$, $j > i$. If $F = e_j e_i^T$, from (5.2)

$$X = \mathrm{up}(e_j e_i^T + \zeta_D e_i e_j^T),$$
$$\|X\|_F = \sqrt{1+\zeta_D^2} = \sqrt{1+\zeta_D^2}\kappa_2(\bar{R})\|F\|_F.$$

So we see that, at least for diagonal R, the bounds are tight, and in this restricted case $\kappa_{ME}(A)$ *is* the condition number.

This refined matrix equation analysis shows to what extent the solution X of (5.1), and so the sensitivity of R in the QR factorization, is dependent on the row scaling in $R = D\bar{R}$. From the term $D^{-1}\mathrm{up}(\bar{R}^{-T}F^T)D$ in (5.2), we saw that multipliers δ_j/δ_i occurred only with $j > i$. As a result we obtained ζ_D in our bounds rather than $\kappa_2(D)$, where

$$\zeta_D \le \kappa_2(D),$$

with equality if and only if the minimum element comes before the maximum on the diagonal. Thus we obtained full cancellation of D^{-1} with D in the first term on the right-hand side of (5.2) and partial cancellation in the second.

This gives some insight as to why R in the QR factorization is less sensitive than the earlier condition estimator $\sqrt{2}\kappa_2(A)$ indicated. If the ill conditioning of R is mostly due to the bad scaling of its rows, then the correct choice of D in $R = D\bar{R}$ can give $\kappa_2(\bar{R})$ very near one. If at the same time ζ_D is not large, then $\kappa(R, D)$ in (5.10) can be much smaller than $\sqrt{2}\kappa_2(R)$; see (5.11). Standard pivoting always ensures that such a D exists, and in fact gives (5.14). However, if we do not use any pivoting, then Remark 5.1 suggests that any relatively small earlier elements on the diagonal of R could correspond to poor conditioning of the factorization.

We will return to $\kappa_{ME}(A)$ and $\kappa(R, D)$ when we seek practical estimates of the condition number that we derive in the next section.

5.2. Matrix–vector equation analysis for *R*. We can now obtain provably sharp, but less intuitive, results by viewing the matrix equation (5.1) 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")

$$\mathrm{uvec}(C) \equiv \begin{bmatrix} c_1^{(1)} \\ c_2^{(2)} \\ \cdot \\ c_n^{(n)} \end{bmatrix}.$$

It is the vector formed by stacking the columns of the upper triangular part of C into one long vector.

To analyze (3.5) we again consider the general form (5.1), repeated here for clarity:

$$R^TX + X^TR = R^TF + F^TR, \quad X \text{ and } R \text{ upper triangular, } R \text{ nonsingular,} \tag{5.15}$$

which we saw via (2.4) has a unique upper triangular solution X. The upper and lower triangular parts of (5.1) contain identical information, and we now write the upper triangular part in matrix–vector, rather than matrix–matrix, format. The first j elements of the jth column of (5.15) are given by

$$R_{jj}^T x_j^{(j)} + X_{jj}^T r_j^{(j)} = R_{jj}^T f_j^{(j)} + \begin{bmatrix} f_1^{(j)T} \\ f_2^{(j)T} \\ \cdot \\ f_j^{(j)T} \end{bmatrix} r_j^{(j)},$$

and by rewriting this, we can see how to solve for $x_j^{(j)}$, $j = 1, \ldots, n$:

$$(R_{jj}^T + e_j r_j^{(j)T}) x_j^{(j)} + \begin{bmatrix} r_j^{(1)T} x_1^{(1)} \\ r_j^{(2)T} x_2^{(2)} \\ \cdot \\ r_j^{(j-1)T} x_{j-1}^{(j-1)} \\ 0 \end{bmatrix} = (R_{jj}^T + e_j r_j^{(j)T}) f_j^{(j)} + \begin{bmatrix} r_j^{(j)T} f_1^{(j)} \\ r_j^{(j)T} f_2^{(j)} \\ \cdot \\ r_j^{(j)T} f_{j-1}^{(j)} \\ 0 \end{bmatrix},$$

which, upon dividing the last row of this by 2, gives

$$W_R \mathrm{uvec}(X) = Z_R \mathrm{vec}(F), \tag{5.16}$$

where $W_R \in \mathcal{R}^{\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}}$ is

$$\left[\begin{array}{ccc|ccc|c|ccccc}
r_{11} & & & & & & & & & & & \\
r_{12} & r_{11} & & & & & & & & & & \\
 & r_{12} & r_{22} & & & & & & & & & \\
\hline
r_{13} & & & r_{11} & & & & & & & & \\
 & r_{13} & r_{23} & r_{12} & r_{22} & & & & & & & \\
 & & & r_{13} & r_{23} & r_{33} & & & & & & \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & & & \\
\hline
r_{1n} & & & & & & & r_{11} & & & & \\
 & r_{1n} & r_{2n} & & & & & r_{12} & r_{22} & & & \\
 & & & r_{1n} & r_{2n} & r_{3n} & & r_{13} & r_{23} & r_{33} & & \\
 & & & & & & \cdot & \cdot & \cdot & \cdot & \cdot & \\
 & & & & & & & r_{1n} & r_{2n} & r_{3n} & \cdot & r_{nn}
\end{array}\right]$$

and $Z_R \in \mathcal{R}^{\frac{n(n+1)}{2} \times n^2}$ is

$$\left[\begin{array}{cccc|cccc|c|cccc}
r_{11} & & & & & & & & & & & \\
\hline
r_{12} & r_{22} & & & r_{11} & & & & & & & \\
 & & & & r_{12} & r_{22} & & & & & & \\
\hline
\cdot & \cdot & \cdot & & \cdot & \cdot & \cdot & & \cdot & & & \\
\hline
r_{1n} & r_{2n} & \cdot & r_{nn} & & & & & & r_{11} & & \\
 & & & & r_{1n} & r_{2n} & \cdot & r_{nn} & & r_{12} & r_{22} & \\
 & & & & & & & & \cdot & \cdot & \cdot & \cdot \\
 & & & & & & & & & r_{1n} & r_{2n} & \cdot & r_{nn}
\end{array}\right].$$

Since R is nonsingular, W_R is also, and from (5.16)

$$\text{uvec}(X) = W_R^{-1} Z_R \text{vec}(F). \tag{5.17}$$

Remembering X is upper triangular, we see that

$$\begin{aligned} \|X\|_F &= \|\text{uvec}(X)\|_2 = \|W_R^{-1} Z_R \text{vec}(F)\|_2 \\ &\le \|W_R^{-1} Z_R\|_2 \|\text{vec}(F)\|_2 = \|W_R^{-1} Z_R\|_2 \|F\|_F, \end{aligned} \tag{5.18}$$

where for *any* nonsingular upper triangular R equality can be obtained by choosing $\text{vec}(F)$ to lie in the space spanned by the right singular vectors corresponding to the largest singular value of $W_R^{-1} Z_R$. It follows that (5.18) is tight, so from (5.8), derived from the matrix equation approach, and (5.11)

$$\|W_R^{-1} Z_R\|_2 \le \kappa_{ME}(A) \le \sqrt{2}\kappa_2(A). \tag{5.19}$$

Remark 5.2. Usually the first and second inequalities are strict. For example, let $R = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. Then we obtain $\|W_R^{-1} Z_R\|_2 = 1.7321$, $\kappa_{ME}(A) = 2.8679$, and $\sqrt{2}\kappa_2(A) = 3.7025$ by MATLAB ($\kappa_{ME}(A) = 2.8679$ was obtained via an optimization problem). But from Remark 5.1 the first inequality becomes an equality if R is diagonal. The second also becomes an equality if R is an $n \times n$ identity matrix with $n \ge 2$.

The structure of W_R and Z_R reveals that each column of W_R is one of the columns of Z_R, and so $W_R^{-1} Z_R$ has an $n(n+1)/2$ square identity submatrix, giving

$$\|W_R^{-1} Z_R\|_2 \ge 1. \tag{5.20}$$

Remark 5.3. We can obtain no better constant lower bound than this, as can be seen by taking $R = \text{diag}(1, \delta, \ldots, \delta^{n-1})$, $0 < \delta \le 1$, for by taking $D = R$ in (5.19), (5.9), and (5.10), we see from Remark 5.1 that

$$1 \le \|W_R^{-1} Z_R\|_2 = \kappa_{ME}(A) = \kappa(R, D) = \sqrt{1 + \delta^2} \to 1 \quad \text{as} \quad \delta \to 0. \tag{5.21}$$

These results, and the analysis in section 4 for Q_1, lead to our new first-order norm-based perturbation theorem.

THEOREM 5.2. *Let* $A = (Q_1, Q_2) \begin{pmatrix} \text{R} \\ 0 \end{pmatrix}$ *be the QR factorization of* $A \in \mathcal{R}^{m \times n}$ *with full column rank, and let* ΔA *be a real* $m \times n$ *matrix. Let* $\epsilon \equiv \|\Delta A\|_F / \|A\|_2$, $\epsilon_1 \equiv \|Q_1^T \Delta A\|_F / \|A\|_2$, *and* $\epsilon_2 \equiv \|Q_2^T \Delta A\|_F / \|A\|_2$. *If* (2.8) *holds, then there is a unique QR factorization satisfying*

$$A + \Delta A = (Q_1 + \Delta Q_1)(R + \Delta R), \tag{5.22}$$

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le \kappa_R(A)\epsilon_1 + O(\epsilon^2), \tag{5.23}$$

where with W_R *and* Z_R *as in* (5.16) *and* $\kappa_{ME}(A)$ *as in* (5.9) *and* (5.10),

$$1 \le \kappa_R(A) \equiv \|W_R^{-1} Z_R\|_2 \le \kappa_{ME}(A) \le \sqrt{2}\kappa_2(A), \tag{5.24}$$

and with

$$\kappa_{Q_1}(A) \equiv \sqrt{2}\|R_{n-1}^{-1}\|_2 \|A\|_2 \le \sqrt{2}\kappa_2(A),$$

$$\|Q_2^T \Delta Q_1\|_F \le \kappa_2(A)\epsilon_2 + O(\epsilon^2),$$
$$\|Q_1^T \Delta Q_1\|_F \le \kappa_{Q_1}(A)\epsilon_1 + O(\epsilon^2).$$

Proof. From Corollary 2.2 $A + \Delta A$ has the unique QR factorization (5.22). From (3.5), (5.15), (5.17) with $G \equiv \Delta A/\epsilon$, and (2.5) we have

$$\mathrm{uvec}(\dot{R}(0)) = W_R^{-1} Z_R \mathrm{vec}(Q_1^T G),$$

so taking the 2-norm gives

$$\|\dot{R}(0)\|_F \le \|W_R^{-1} Z_R\|_2 \|Q_1^T G\|_F = \|W_R^{-1} Z_R\|_2 \|A\|_2 \epsilon_1/\epsilon.$$

With $\|A\|_2 = \|R\|_2$, $\|A^\dagger\|_2 = \|R^{-1}\|_2$, (5.19), and (5.20), this gives (5.24) and

$$\frac{\|\dot{R}(0)\|_F}{\|R\|_2} \le \kappa_R(A)\epsilon_1/\epsilon. \tag{5.25}$$

Thus, from the Taylor series (3.9) of $R(t)$, (5.23) follows. The remaining results are restatements of (5.20), (5.19), (4.1), and (4.3). □

Remark 5.4. From (5.24) we know that the first-order perturbation bound (5.23) is at least as good as (3.2). In fact, it can be better by an arbitrary factor. Consider the example in Remark 5.3, where taking $D = R$,

$$\kappa_R(A) = \kappa_{ME}(A) = \kappa(R, D) = \sqrt{1 + \delta^2}, \quad \kappa_2(A) = 1/\delta$$

and

$$\frac{\sqrt{2}\kappa_2(A)}{\kappa_R(A)} \sim \frac{\sqrt{2}}{\delta} \quad \text{as } \delta \to 0.$$

We see that the first-order perturbation bound (3.2) can severely overestimate the effect of a perturbation in A.

Remark 5.5. If we take $R = \mathrm{diag}(\delta^{1-n}, \ldots, \delta, 1)$, $0 < \delta \le 1$, we see that $\kappa_2(R) = \kappa_2(A) = \delta^{1-n}$, while

$$\kappa_R(A) = \kappa_{ME}(A) = \kappa(R, D) = \sqrt{1 + \delta^{2-2n}},$$

which is close to the upper bound $\sqrt{2}\kappa_2(A)$ for small δ. This shows that relatively small early diagonal elements of R cause poor condition and suggests that if we do not use pivoting, then there is a significant chance that the condition of the problem will approach its upper bound, at least for randomly chosen matrices.

When we use standard *pivoting*, we see from (5.24) and (5.14) that

$$1 \le \kappa_R(AP) \equiv \|W_R^{-1} Z_R\|_2 \le \kappa_{ME}(AP) \le 2^{n-1}\sqrt{n(n+1)},$$

but the following tighter result is shown in [2, Theorem 2.2].

THEOREM 5.3. *Let $A \in \mathcal{R}^{m\times n}$ be of full column rank, with the QR factorization $AP = Q_1 R$ when the standard column pivoting strategy is used. Then*

$$1 \le \kappa_R(AP) = \|W_R^{-1} Z_R\|_2 \le \|W_R^{-1} Z_R\|_F \le \sqrt{\frac{1}{27}4^{n+1} + \frac{1}{3}n^2 + \frac{2}{9}n - \frac{4}{27}}. \tag{5.26}$$

There is a parametrized family of matrices $A(\theta)$, $\theta \in (0, \pi/2]$, for which

$$\|W_R^{-1} Z_R\|_F \to \sqrt{\frac{1}{27}4^{n+1} + \frac{1}{3}n^2 + \frac{2}{9}n - \frac{4}{27}} \quad \textit{as} \quad \theta \to 0. \qquad \square$$

Theorem 5.3 shows that when the standard column pivoting strategy is used, $\kappa_R(AP)$ is bounded for fixed n no matter how large $\kappa_2(A)$ is. Many numerical experiments with this strategy suggest that $\kappa_R(AP)$ is usually close to its lower bound of 1, but we give an extreme example in section 6 where it is not.

When we do not use pivoting, we have no such simple result for $\kappa_R(A)$, and it is, as far as we can see, unreasonably expensive to compute or approximate $\kappa_R(A)$ directly with the usual approach. Fortunately, $\kappa_{ME}(A)$ is apparently an excellent approximation to $\kappa_R(A)$, and $\kappa_{ME}(A)$ *is* quite easy to estimate. All we need to do is choose a suitable D in $\kappa(R, D)$ in (5.10). We now consider how to do this.

6. Numerical experiments and condition estimators. In section 5 we presented new first-order perturbation bounds for the R factor of the QR factorization using two different approaches, defined $\kappa_R(A) \equiv \|W_R^{-1} Z_R\|_2$ as the condition number for the R factor, and suggested $\kappa_R(A)$ could be estimated in practice by $\kappa(R, D)$. Our new first-order results are sharper than previous results for R and at least as sharp for Q_1, and we give some numerical tests to illustrate both this and possible estimators for $\kappa_R(A)$.

We would like to choose D such that $\kappa(R, D)$ is a good approximation to the minimum $\kappa_{ME}(A)$ in (5.9) and show that this is a good estimate of the condition number $\kappa_R(A)$. Then a procedure for obtaining an $O(n^2)$ condition estimator for R in the QR factorization (i.e., an estimator for $\kappa_R(A)$) is to choose such a D, use a standard condition estimator (see, for example, [6]) to estimate $\kappa_2(D^{-1}R)$, and take $\kappa(R, D)$ in (5.10) as the appropriate estimate.

By a well-known result of van der Sluis [18], $\kappa_2(D^{-1}R)$ will be nearly minimal when the rows of $D^{-1}R$ are equilibrated. But this could lead to a large ζ_D in (5.10). There are three obvious possibilities for D. The first one is choosing D to equilibrate R precisely. Specifically, take $\delta_i = \sqrt{\sum_{j=i}^n r_{ij}^2}$ for $i = 1, \ldots, n$. The second one is choosing D to equilibrate R as far as possible while keeping $\zeta_D \le 1$. Specifically, take $\delta_1 = \sqrt{\sum_{j=1}^n r_{1j}^2}$, $\delta_i = \sqrt{\sum_{j=i}^n r_{ij}^2}$ if $\sqrt{\sum_{j=i}^n r_{ij}^2} \le \delta_{i-1}$; otherwise $\delta_i = \delta_{i-1}$ for $i = 2, \ldots, n$. The third one is choosing $\delta_i = r_{ii}$. Computations show that the third choice can sometimes cause unnecessarily large estimates, so we will not give any results for that choice. We specify the diagonal matrix D obtained by the first method and the second method by D_1 and D_2, respectively, in the following.

We give three sets of examples. The first set of matrices are $n \times n$ Pascal matrices (with elements $a_{1j} = a_{i1} = 1$, $a_{ij} = a_{i,j-1} + a_{i-1,j}$), $n = 1, 2, \ldots, 15$. The results are shown in Table 6.1 without pivoting, giving $A = Q_1 R$, and in Table 6.2 with pivoting, giving $AP = \tilde{Q}_1 \tilde{R}$. Note in Table 6.1 how the upper bound $\sqrt{2}\kappa_2(A)$ can be far worse than the condition number $\kappa_R(A)$, which itself can be much greater than its lower bound of 1. In Table 6.2 pivoting is seen to give a significant improvement to $\kappa_R(A)$, bringing $\kappa_R(AP)$ very close to its lower bound, but of course $\sqrt{2}\kappa_2(AP) = \sqrt{2}\kappa_2(A)$ still. Also, we observe from Table 6.1 that both $\kappa(R, D_1)$ and $\kappa(R, D_2)$ are very good estimates for $\kappa_R(A)$. The latter is a little better than the former. In Table 6.2 $\kappa(\tilde{R}, D_1) = \kappa(\tilde{R}, D_2)$ (in fact $D_1 = D_2$), and they are also good estimates for $\kappa_R(AP)$.

TABLE 6.1
Results for Pascal matrices without pivoting, $A = Q_1 R$.

n	$\kappa_R(A)$	$\kappa(R, D1)$	$\kappa(R, D2)$	$\kappa_{Q_1}(A)$	$\sqrt{2}\kappa_2(A)$
1	1.0e+00	1.4e+00	1.4e+00	—	1.4e+00
2	1.9e+00	3.4e+00	1.9e+00	2.6e+00	9.7e+00
3	4.6e+00	1.4e+01	1.4e+01	1.9e+01	8.8e+01
4	1.4e+01	6.1e+01	6.1e+01	1.6e+02	9.8e+02
5	5.0e+01	2.6e+02	2.6e+02	1.6e+03	1.2e+04
6	1.8e+02	1.1e+03	1.1e+03	1.8e+04	1.6e+05
7	6.7e+02	4.5e+03	4.2e+03	2.2e+05	2.1e+06
8	2.5e+03	1.8e+04	1.7e+04	2.8e+06	2.9e+07
9	9.4e+03	7.4e+04	6.6e+04	3.6e+07	4.1e+08
10	3.6e+04	3.0e+05	2.6e+05	4.8e+08	5.9e+09
11	1.4e+05	1.2e+06	1.1e+06	6.6e+09	8.5e+10
12	5.2e+05	4.9e+06	4.2e+06	9.1e+10	1.2e+12
13	2.0e+06	2.0e+07	1.7e+07	1.3e+12	1.8e+13
14	7.8e+06	8.0e+07	6.6e+07	1.8e+13	2.7e+14
15	3.0e+07	3.2e+08	2.6e+08	2.6e+14	4.0e+15

TABLE 6.2
Results for Pascal matrices with pivoting, $AP = \tilde{Q}_1 \tilde{R}$.

n	$\kappa_R(AP)$	$\kappa(\tilde{R}, D1)$	$\kappa(\tilde{R}, D2)$	$\kappa_{Q_1}(AP)$	$\sqrt{2}\kappa_2(A)$
1	1.0e+00	1.4e+00	1.4e+00	—	1.4e+00
2	1.2e+00	1.8e+00	1.8e+00	1.7e+00	9.7e+00
3	1.3e+00	2.2e+00	2.2e+00	1.3e+01	8.8e+01
4	1.7e+00	3.4e+00	3.4e+00	1.1e+02	9.8e+02
5	1.8e+00	4.1e+00	4.1e+00	1.0e+03	1.2e+04
6	2.2e+00	4.7e+00	4.7e+00	7.5e+03	1.6e+05
7	2.1e+00	5.1e+00	5.1e+00	8.5e+04	2.1e+06
8	2.6e+00	6.5e+00	6.5e+00	1.2e+06	2.9e+07
9	3.5e+00	8.8e+00	8.8e+00	1.5e+07	4.1e+08
10	3.4e+00	9.4e+00	9.4e+00	2.4e+08	5.9e+09
11	3.4e+00	9.2e+00	9.2e+00	2.3e+09	8.5e+10
12	3.3e+00	9.7e+00	9.7e+00	3.0e+10	1.2e+12
13	3.3e+00	1.1e+01	1.1e+01	3.5e+11	1.8e+13
14	3.6e+00	1.2e+01	1.2e+01	5.4e+12	2.7e+14
15	3.3e+00	1.2e+01	1.2e+01	8.6e+13	4.0e+15

The second set of matrices are 10×8 A_j, $j = 1, 2, \ldots, 8$, which are all obtained from the same random 10×8 matrix (produced by the MATLAB function randn), but with its jth column multiplied by 10^{-8} to give A_j. The results without pivoting are shown in Table 6.3. All the results with pivoting are similar to that for $j = 8$ in Table 6.3, and so are not given here. For $j = 1, 2 \ldots, 7$, $\kappa_R(A)$ and $\kappa_{Q_1}(A)$ are both close to their upper bound $\sqrt{2}\kappa_2(A)$, but for $j = 8$, both $\kappa_R(A)$ and $\kappa_{Q_1}(A)$ are significantly smaller than $\sqrt{2}\kappa_2(A)$. All these results are what we expected, since the matrix R is ill conditioned because r_{jj} is very small, but for $j = 1, 2, \ldots, 7$ the rows of R are already essentially equilibrated, and we do not expect $\kappa_R(A)$ to be much better than $\sqrt{2}\kappa_2(A)$. Also, for the first seven cases the smallest singular value of the leading part R_{n-1} is close to that of R, so that $\kappa_{Q_1}(A)$ could not be much better than $\sqrt{2}\kappa_2(A)$. For $j = 8$, even though R is still ill conditioned because $r_{8,8}$ is very small, it is not at all equilibrated and becomes well conditioned by row scaling. Notice at the same time that ζ_D is close to 1, so $\kappa(R, D1)$, $\kappa(R, D_2)$, and therefore $\kappa_R(A)$ are much better than $\sqrt{2}\kappa_2(A)$. In this case, the smallest singular value of R is significantly smaller than that of R_{n-1}. Thus $\kappa_{Q_1}(A)$, the condition number for the change in Q_1 lying in the range of Q_1, is spectacularly better than $\sqrt{2}\kappa_2(A)$. This is a contrived example,

but it serves to emphasize the benefits of pivoting for the condition of both Q_1 and R.

TABLE 6.3
Results for 10×8 matrix A_j, $j = 1, \ldots, 8$, without pivoting.

j	$\kappa_R(A)$	$\kappa(R, D1)$	$\kappa(R, D2)$	$\kappa_{Q_1}(A)$	$\sqrt{2}\kappa_2(A)$
1	1.9e+08	4.0e+08	3.0e+08	3.0e+08	4.8e+08
2	1.3e+08	2.9e+08	2.7e+08	2.6e+08	3.8e+08
3	1.9e+08	4.5e+08	3.9e+08	4.7e+08	5.5e+08
4	1.4e+08	3.1e+08	2.6e+08	2.9e+08	4.5e+08
5	1.2e+08	3.1e+08	2.4e+08	3.9e+08	4.2e+08
6	8.8e+07	2.2e+08	1.7e+08	3.5e+08	3.9e+08
7	9.3e+07	2.1e+08	1.7e+08	4.4e+08	5.5e+08
8	2.3e+00	5.5e+00	4.9e+00	6.6e+00	6.2e+08

The third set of matrices is $n \times n$ upper triangular

$$R = \operatorname{diag}(1, s, \ldots, s^{n-1}) \begin{pmatrix} 1 & -c & -c & \cdot & -c \\ & 1 & -c & \cdot & -c \\ & & 1 & \cdot & -c \\ & & & \cdot & \cdot \\ & & & & 1 \end{pmatrix},$$

where $c = \cos(\theta)$, $s = \sin(\theta)$. These matrices were introduced by Kahan [9]. Of course $Q_1 = I$ here, but the condition numbers depend on R only, and these are all we are interested in. The results for $n = 5, 10, 15, 20, 25$ with $\theta = \pi/8$ are shown in Table 6.4. Again we found $D_1 = D_2$ and list only the results corresponding to D_1.

TABLE 6.4
Results for Kahan matrices, $\theta = \pi/8$.

n	$\kappa_R(A)$	$\kappa(R, D1)$	$\kappa_{Q_1}(A)$	$\sqrt{2}\kappa_2(A)$
5	8.0e+00	1.7e+01	2.2e+02	1.1e+03
10	2.1e+02	6.1e+02	1.0e+06	5.1e+06
15	5.5e+03	2.1e+04	4.0e+09	2.0e+10
20	1.5e+05	6.5e+05	1.5e+13	7.5e+13
25	4.3e+06	2.0e+07	5.4e+16	2.7e+17

In all these examples we see that $\kappa(R, D_1)$ and $\kappa(R, D_2)$ gave excellent estimates for $\kappa_R(A)$, with $\kappa(R, D_2)$ being marginally preferable. For the Kahan matrices, which correspond to correctly pivoted A, we see that in extreme cases, with large enough n, $\kappa_R(A)$ can be large even with standard pivoting. This is about as bad a result as we can get with standard pivoting (it gets a bit worse as $\theta \to 0$ in R), since the Kahan matrices are the parameterized family mentioned in Theorem 5.3. Nevertheless, $\kappa(R, D_1)$ and $\kappa(R, D_2)$ still estimate $\kappa_R(A)$ excellently.

7. Summary and conclusions. The first-order perturbation analyses presented here show just what the sensitivity (condition) of both Q_1 and R is in the QR factorization of full column rank A, and in so doing provide their condition numbers (with respect to the measures used and for sufficiently small ΔA), as well as efficient ways of approximating these. The key norm-based condition numbers we derived for $A + \Delta A = (Q_1 + \Delta Q_1)(R + \Delta R)$ are as follows:

- $\kappa_2(A)$ for that part of ΔQ_1 in $\mathcal{R}(A)^\perp$ (see (4.1)),
- $\kappa_{Q_1}(A) \equiv \sqrt{2}\|R_{n-1}^{-1}\|_2\|A\|_2$ for that part of ΔQ_1 in $\mathcal{R}(A)$ (see (4.3)),
- $\kappa_R(A) \equiv \|W_R^{-1} Z_R\|_2$ for R (see Theorem 5.2),

- the estimate for $\kappa_R(A)$, that is, $\kappa_{ME}(A) \equiv \inf_{D\in\mathcal{D}_n} \kappa(R,D)$, where $\kappa(R,D) \equiv \sqrt{1+\zeta_D^2}\kappa_2(D^{-1}R)$ (see (5.13)).

The condition numbers obey

$$\sqrt{2}\|R_{n-1}^{-1}\|_2\|A\|_2 \le \sqrt{2}\kappa_2(A)$$

for Q_1, while for R

$$1 \le \kappa_R(A) \equiv \|W_R^{-1}Z_R\|_2 \le \kappa_{ME}(A) \le \sqrt{2}\kappa_2(A);$$

see (5.24). The numerical examples and an analysis of the $n = 2$ case (not given here) suggest that $\kappa(R,D)$, with D chosen to equilibrate $\bar{R} \equiv D^{-1}R$ subject to $\zeta_D \le 1$, gives an excellent approximation to $\kappa_R(A)$ in the general case. In the special case of A with orthogonal columns, so R is diagonal, Remark 5.1 showed that by taking $D = R$

$$\kappa_R(A) = \kappa_{ME}(A) = \kappa(R,D) = \sqrt{1+\zeta_D^2} \le \sqrt{2}\kappa_2(D) = \sqrt{2}\kappa_2(A).$$

For general A when we use the standard column pivoting strategy in the QR factorization, $AP = Q_1R$, we saw from (5.14) and [2] that

$$\kappa_{ME}(AP) \le 2^{n-1}\sqrt{n(n+1)},$$
$$\kappa_R(AP) \le \sqrt{\frac{1}{27}4^{n+1} + \frac{1}{3}n^2 + \frac{2}{9}n - \frac{4}{27}}.$$

As a result of these analyses we see that both R and in a certain sense Q_1 can be less sensitive than was thought from previous analyses. The condition numbers depend on any column pivoting of A and show that the standard pivoting strategy often results in a much less sensitive R, and sometimes leads to a much smaller possible change of Q_1 in the range of Q_1, for a given size of perturbation in A.

The matrix equation analysis of section 5.1 also provides a nice analysis of an interesting and possibly more general matrix equation (5.1). The approaches used here are not restricted to this particular analysis, but are powerful general tools and appear to be applicable to any matrix factorization.

All of the new bounds here are first-order bounds. They are *asymptotically* correct, but it is possible that for $\|Q_1^T\Delta Q_1\|_F$ the second-order term in (4.3) will blow up when ϵ is significant. This is addressed in [3]. For $\|Q_2^T\Delta Q_1\|_F$, the second-order term cannot blow up under condition (2.8) by the result of Sun [17]. For $\|\Delta R\|_F/\|R\|_2$, according to the result of Stewart [11], the second-order term in (5.23) will not blow up when

$$\|R^{-T}(A^T\Delta A + \Delta A^T A + \Delta A^T\Delta A)R^{-1}\|_F < 1/2.$$

By following the approach of Stewart [10, Theorem 3.1] (see also [14, Theorem 2.11]), it would be straightforward, but detailed and lengthy, to extend our first-order results to provide strict perturbation bounds, as was done in [4]. We could also provide new bounds on the components of ΔR, but we chose not to do either of these here in order to keep the material and the basic ideas as brief and approachable as possible. Our condition numbers and resulting bounds are asymptotically sharp, so there is less need for strict bounds. A new bound on the components of ΔR is given in [2].

Acknowledgments. We would like to thank Ji-guang Sun for his suggestions and encouragement and for providing us with draft versions of his work, and Nick Higham for his hospitality, assistance, and clear comments.

REFERENCES

[1] X.-W. Chang, *Perturbation Analysis of Some Matrix Factorizations*, Ph.D. thesis, McGill University, School of Computer Science, Montreal, Quebec, Canada, 1997, in preparation.
[2] X.-W. Chang and C. C. Paige, *A Perturbation Analysis for R in the QR Factorization,* Technical report SOCS-95.7, McGill University, School of Computer Science, Montreal, Quebec, Canada, 1995.
[3] X.-W. Chang and C. C. Paige, *Perturbation analyses for the QR factorization with bounds on changes in the elements of A*, manuscript.
[4] 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.
[5] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[6] N. J. Higham, *A survey of condition number estimation for triangular matrices*, SIAM Rev., 29 (1987), pp. 575–596
[7] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1996.
[8] Y. P. Hong and C.-T. Pan, *Rank-revealing QR factorizations and the singular value decomposition*, Math. Comp., 58 (1992), pp. 213–232.
[9] W. Kahan, *Numerical linear algebra*, Canad. Math. Bull., 9 (1966), pp. 757–801.
[10] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.
[11] G. W. Stewart, *Perturbation bounds for the QR factorization of a matrix*, SIAM J. Numer. Anal., 14 (1977), pp. 509–518.
[12] G. W. Stewart, *On the perturbation of LU, Cholesky, and QR factorizations*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1141–1146.
[13] G. W. Stewart, *On the Perturbation of LU and Cholesky Factors*, Technical report CS-TR-3535 UMIACS-TR-95-93, University of Maryland, Computer Science, College Park, MD, 1995.
[14] G W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, San Diego, 1990.
[15] J.-G. Sun, *Perturbation bounds for the Cholesky and QR factorization*, BIT, 31 (1991), pp. 341–352.
[16] J.-G. Sun, *Componentwise perturbation bounds for some matrix decompositions*, BIT, 32 (1992), pp. 702–714.
[17] J.-G. Sun, *On perturbation bounds for the QR factorization*, Linear Algebra Appl., 215 (1995), pp. 95–112.
[18] A. van der Sluis, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969), pp. 14–23.
[19] H. Zha, *A componentwise perturbation analysis of the QR decomposition*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1124–1131.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 4, pp. 793–817, October 1997

001

ON THE LIDSKII–VISHIK–LYUSTERNIK PERTURBATION THEORY FOR EIGENVALUES OF MATRICES WITH ARBITRARY JORDAN STRUCTURE*

JULIO MORO†, JAMES V. BURKE‡, AND MICHAEL L. OVERTON§

Dedicated to V. B. Lidskii and M. I. Vishik on the respective occasions of their 70th and 75th birthdays.

Abstract. Let A be a complex matrix with arbitrary Jordan structure and λ an eigenvalue of A whose largest Jordan block has size n. We review previous results due to Lidskii [*U.S.S.R. Comput. Math. and Math. Phys.*, 1 (1965), pp. 73–85], showing that the splitting of λ under a small perturbation of A of order ε is, generically, of order $\varepsilon^{1/n}$. Explicit formulas for the leading coefficients are obtained, involving the perturbation matrix and the eigenvectors of A. We also present an alternative proof of Lidskii's main theorem, based on the use of the Newton diagram. This approach clarifies certain difficulties which arise in the nongeneric case and leads, in some situations, to the extension of Lidskii's results. These results suggest a new notion of Hölder condition number for multiple eigenvalues, depending only on the associated left and right eigenvectors, appropriately normalized, not on the Jordan vectors.

Key words. perturbation of eigenvalues, perturbation theory for linear operators, stability theory, Newton diagram, Newton envelope, spectral condition number

AMS subject classifications. 15A18, 34D10, 47A55, 65F15

PII. S0895479895294666

1. Introduction.

Given a square complex matrix A, it is an important question from both the theoretical and the practical points of view to know how the eigenvalues and eigenvectors change when the elements of A are subjected to small perturbations. The usual formulation of the problem introduces a perturbation parameter ε, belonging to some neighborhood of zero, and writes the perturbed matrix as $A + \varepsilon B$ for an arbitrary matrix B. In this situation, it is well known [1, section 9.3.1], [7, section II.1.2] that each eigenvalue or eigenvector of $A + \varepsilon B$ admits an expansion in fractional powers of ε, whose zero-th order term is an eigenvalue or eigenvector of the unperturbed matrix A.

In this paper we address the question of determining the first-order term of this expansion or, more precisely, the first nonzero perturbation term. No restriction is imposed on the Jordan structure of A, although we assume that this Jordan structure is known from the outset. In section 2 we present two results, due to Lidskii [10], which provide, under certain nondegeneracy conditions, the leading exponents and leading coefficients of both eigenvalue and eigenvector perturbations. The central idea of the

*Received by the editors November 9, 1995; accepted for publication (in revised form) by B. Kågström August 30, 1996.

http://www.siam.org/journals/simax/18-4/29466.html

†Departamento de Matemáticas, Universidad Carlos III, Madrid, Spain (jmoro@dulcinea.uc3m.es). This work was done while the author was on leave from Universidad Complutense de Madrid, visiting the Courant Institute of Mathematical Sciences, New York University, New York, NY. The work of this author was supported in part by an Ayuda Complutense Postdoctoral en el Extranjero and by CICYT grant TAP94-115.

‡Department of Mathematics, University of Washington, Seattle, WA 98195-4350 (burke@math.washington.edu). The work of this author was supported in part by NSF grant DMS-9303772.

§Computer Science Department, Courant Institute of Mathematical Sciences, New York University, New York, NY (overton@cs.nyu.edu). The work of this author was supported in part by NSF grant CCR-9401136.

proof is simply to transform the characteristic equation $\det(\omega I - A - \varepsilon B) = 0$ into an equivalent one $Q(\mu, z) = 0$ through a change of variables

$$z = \varepsilon^{1/n},$$

$$\mu = \frac{\omega - \lambda}{z}$$

for a suitable n, where λ is an eigenvalue of A. An appropriate factorization of $Q(\mu, 0)$ leads to the final result.

A specific example may be helpful to give a better idea of these results: take a 9×9 Jordan matrix J with a unique zero eigenvalue and four Jordan blocks with respective dimensions 3, 3, 2, and 1. Lidskii's results show that, given a small perturbation $J + \varepsilon B$, every Jordan block of J of dimension n gives rise, generically, to n eigenvalues of the perturbed matrix with leading term $O(\varepsilon^{1/n})$. In this particular case, this amounts to six eigenvalues of order $\varepsilon^{1/3}$, two of order $\varepsilon^{1/2}$, and one of order ε. As for the coefficients of these leading terms, we will show that they depend *exclusively* on the elements of B marked with a box in the matrix below:

$$B = \left[\begin{array}{ccc|ccc|ccc} * & * & * & * & * & * & * & * & * \\ * & * & * & * & * & * & * & * & * \\ \boxed{} & * & * & \boxed{} & * & * & \boxed{} & * & \boxed{} \\ \hline * & * & * & * & * & * & * & * & * \\ * & * & * & * & * & * & * & * & * \\ \boxed{} & * & * & \boxed{} & * & * & \boxed{} & * & \boxed{} \\ \hline * & * & * & * & * & * & * & * & * \\ \boxed{} & * & * & \boxed{} & * & * & \boxed{} & * & \boxed{} \\ \hline \boxed{} & * & * & \boxed{} & * & * & \boxed{} & * & \boxed{} \end{array}\right].$$

More specifically, let Φ_1 denote the 2×2 matrix given by the four boxes at the top left, i.e.,

$$\Phi_1 = \begin{bmatrix} B_{31} & B_{34} \\ B_{61} & B_{64} \end{bmatrix},$$

and let ξ_1^1, ξ_1^2 be the eigenvalues of Φ_1. Then the perturbed matrix $J + \varepsilon B$ has six eigenvalues with leading terms

$$(\xi_1^k)^{1/3} \varepsilon^{1/3}, \qquad k = 1, 2,$$

using all three cube roots of each ξ_1^k. Now, let

$$\Phi_2 = \left[\begin{array}{cc|c} B_{31} & B_{34} & B_{37} \\ B_{61} & B_{64} & B_{67} \\ \hline B_{81} & B_{84} & B_{87} \end{array}\right]$$

and let ξ_2 denote the *Schur complement of* Φ_1 *in* Φ_2, i.e.,

$$\xi_2 = B_{87} - [B_{81}\ B_{84}]\, \Phi_1^{-1} \begin{bmatrix} B_{37} \\ B_{67} \end{bmatrix}.$$

Then the two $O(\varepsilon^{1/2})$ eigenvalues of $J + \varepsilon B$ have leading terms

$$(\xi_2)^{1/2}\varepsilon^{1/2}.$$

Finally, the leading coefficient of the $O(\varepsilon)$ eigenvalue is the Schur complement of Φ_2 in the 9×9 matrix formed by all boxes, i.e.,

$$\xi_3 = B_{99} - [B_{91}\ B_{94}\ B_{97}]\,\Phi_2^{-1} \begin{bmatrix} B_{39} \\ B_{69} \\ B_{89} \end{bmatrix}.$$

In the most general case when A is not in Jordan form, one must replace the elements of B marked with the boxes by products yBx, where x (resp., y) is a right (resp., left) eigenvector of A.

The first results in this direction were obtained by Vishik and Lyusternik [13], motivated by applications to differential operators. Lidskii [10] generalized their results in the finite-dimensional case, obtaining simple explicit formulas for the perturbation coefficients and providing, at the same time, a much more elementary proof (which is essentially the one we present in section 2). The results in both [13] and [10] were later refined by Baumgärtel [1, section 7.4] in the sense of dealing not only with perturbation series for eigenvalues and eigenvectors, but also with the corresponding eigenprojections as functions of ε. Vainberg and Trenogin [12, section 32], on the other hand, offer a fairly thorough account of similar results, obtained for Fredholm operators by applying the techniques of branching theory. Langer and Najman [9] recently generalized Lidskii's results to matrix pencils $M(\lambda) + N(\lambda, \varepsilon)$, using the local Smith normal form of parameter-dependent matrices (Lidskii's results follow from choosing $M \equiv A - \lambda I$, $N \equiv \varepsilon B$). The fundamental results of Lidskii remain, however, almost completely unknown in the Western literature. The only references to [10] appearing in the Science Citation Index are [3] and [9], and both of these continue earlier work [2], [8] in which the authors were unaware of [10]. The main purpose of this paper is, therefore, to bring Lidskii's results to the attention of the broad linear algebra community. See [11] for an application to stability theory for Hamiltonian systems.

Section 2 is devoted to reviewing both the results and the proofs given in [10]. We should stress here that, although Lidskii stated his results as being valid for analytic perturbations, we will see that they hold in fact for a more general class of perturbations, including those of class C^1 (see Remark 4 in section 2). Lidskii's results, however, depend on certain nondegeneracy assumptions, and no information about the leading exponents or coefficients is available in the degenerate case from the approach taken in section 2. Consider, for instance, the following example taken from Wilkinson [14, section 2.22]: let A be a Jordan matrix with two Jordan blocks of sizes 3 and 2, which is perturbed only in the positions (3,4) and (5,1), i.e.,

$$A + \varepsilon B = \left[\begin{array}{ccc|cc} 0 & 1 & & & \\ & 0 & 1 & & \\ & & 0 & \varepsilon & \\ \hline & & & 0 & 1 \\ \varepsilon & & & & 0 \end{array}\right]. \tag{1.1}$$

One can easily check that the characteristic polynomial of $A + \varepsilon B$ is $\varepsilon^2 - \lambda^5$. Hence, the eigenvalues of $A + \varepsilon B$ are $O(\varepsilon^{2/5})$, an order which Lidskii's results are unable to

predict. We present in section 3 a different approach which will, in particular, reveal the origin of this exponent. Apart from providing an alternative proof of Lidskii's main theorem, the new point of view identifies the difficulties which arise in the degenerate case and, in some situations, leads to extensions of the results in section 2. This alternative approach is much in the spirit of [12] since our main tool is the Newton diagram.

We end by proposing in section 4 a new notion of Hölder condition number for multiple eigenvalues, suggested by Lidskii's results. Although it is closely related to previous Hölder condition numbers in the literature [4, p. 156] its main difference is that it depends only on the associated left and right eigenvectors, appropriately normalized, not on the Jordan vectors.

2. Lidskii's perturbation theory. Let A be a complex matrix with Jordan form

$$\left[\begin{array}{c|c} J & \\ \hline & \widehat{J} \end{array}\right] = \left[\begin{array}{c} Q \\ \hline \widehat{Q} \end{array}\right] A \left[\begin{array}{c|c} P & \widehat{P} \end{array}\right] \tag{2.1}$$

with

$$\left[\begin{array}{c} Q \\ \hline \widehat{Q} \end{array}\right] \left[\begin{array}{c|c} P & \widehat{P} \end{array}\right] = I, \tag{2.2}$$

where J corresponds to a multiple eigenvalue λ and $\widehat{J}$ is the part of the Jordan form containing the other eigenvalues of A. Let

$$J = \mathrm{Diag}(\Gamma_1^1, \ldots, \Gamma_1^{r_1}, \ldots, \Gamma_q^1, \ldots, \Gamma_q^{r_q}), \tag{2.3}$$

where, for $j = 1, \ldots, q$,

$$\Gamma_j^1 = \cdots = \Gamma_j^{r_j} = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix}$$

is a Jordan block of dimension n_j, repeated r_j times, and ordered so that

$$n_1 > n_2 > \cdots > n_q.$$

The n_j are called the *partial multiplicities* for λ. The eigenvalue λ is semisimple (nondefective) if $q = n_1 = 1$ and nonderogatory if $q = r_1 = 1$. The algebraic and geometric multiplicities of λ are, respectively,

$$m = \sum_{j=1}^{q} r_j n_j \quad \text{and} \quad g = \sum_{j=1}^{q} r_j.$$

We further partition

$$P = \left[\begin{array}{c|c|c|c|c|c|c} & & & & & & \\ P_1^1 & \dots & P_1^{r_1} & \dots & P_q^1 & \dots & P_q^{r_q} \\ & & & & & & \end{array}\right]$$

conformally with (2.3). The columns of each P_j^k form a right Jordan chain of A with length n_j corresponding to λ. If we denote by x_j^k the first column of P_j^k, each x_j^k is a right eigenvector of A associated with λ. Analogously, we split

$$Q = \left[\begin{array}{c} Q_1^1 \\ \hline \vdots \\ \hline Q_1^{r_1} \\ \hline \vdots \\ \hline Q_q^1 \\ \hline \vdots \\ \hline Q_q^{r_q} \end{array}\right]$$

also conformally with (2.3). The rows of each Q_j^k form a left Jordan chain of A of length n_j corresponding to λ. Hence, if we denote by y_j^k the last (i.e., n_jth) row of Q_j^k, each y_j^k is a left eigenvector corresponding to λ. With these eigenvectors we build up matrices

$$Y_j = \left[\begin{array}{c} y_j^1 \\ \vdots \\ y_j^{r_j} \end{array}\right], \qquad X_j = [x_j^1, \dots, x_j^{r_j}],$$

for $j = 1, \dots, q$,

$$W_s = \left[\begin{array}{c} Y_1 \\ \vdots \\ Y_s \end{array}\right], \qquad Z_s = [X_1, \dots, X_s],$$

for $s = 1, \dots, q$, and define square matrices Φ_s and E_s of dimension $f_s = \sum_{j=1}^{s} r_j$ by

$$\Phi_s = W_s B Z_s, \qquad s = 1, \dots, q,$$

$$E_1 = I, \qquad E_s = \left[\begin{array}{cc} 0 & 0 \\ 0 & I \end{array}\right] \qquad \text{for } s = 2, \dots, q,$$

where the identity block in E_s has dimension r_s. Note that, due to the cumulative definitions of W_s and Z_s, every Φ_{s-1}, $s = 2, \dots, q$, is the upper left block of Φ_s.

THEOREM 2.1 (due to Lidskii [10]). *Let $j \in \{1,\dots,q\}$ be given and assume that, if $j>1$, Φ_{j-1} is nonsingular. Then there are $r_j n_j$ eigenvalues of the perturbed matrix $A+\varepsilon B$ admitting a first-order expansion*

$$\lambda_j^{kl}(\varepsilon) = \lambda + (\xi_j^k)^{1/n_j}\,\varepsilon^{1/n_j} + o(\varepsilon^{1/n_j}) \tag{2.4}$$

for $k=1,\dots,r_j$, $l=1,\dots,n_j$, where

(i) *the ξ_j^k, $k=1,\dots,r_j$, are the roots of equation*

$$\det(\Phi_j - \xi\, E_j) = 0 \tag{2.5}$$

or, equivalently, the eigenvalues of the Schur complement of Φ_{j-1} in Φ_j (if $j=1$, the ξ_1^k are just the r_1 eigenvalues of Φ_1),

(ii) *the different values $\lambda_j^{kl}(\varepsilon)$ for $l=1,\dots,n_j$ are defined by taking the n_j distinct n_jth roots of ξ_j^k.*

If, in addition, the r_j solutions ξ_j^k of (2.5) are all distinct, then the eigenvalues (2.4) can be expanded locally in a power series of the form

$$\lambda_j^{kl}(\varepsilon) = \lambda + (\xi_j^k)^{1/n_j}\,\varepsilon^{1/n_j} + \sum_{s=2}^{\infty} a_{js}^{kl}\,\varepsilon^{s/n_j}, \tag{2.6}$$

$k=1,\dots,r_j$, $l=1,\dots,n_j$.

Remark. Two special cases of Theorem 2.1 are well known. In the case in which λ is semisimple, i.e., $q=n_1=1$ with multiplicity r_1, equation (2.4) reduces to

$$\lambda_1^{k1}(\varepsilon) = \lambda + \xi_1^k\,\varepsilon + o(\varepsilon),$$

where the ξ_1^k are the eigenvalues of the $r_1\times r_1$ matrix Y_1BX_1 (cf. [7, section II.2.3]). In the case in which λ is nonderogatory, i.e., $q=r_1=1$ with multiplicity n_1, equation (2.4) reduces to

$$\lambda_j^{1l}(\varepsilon) = \lambda + (\xi_1^1)^{1/n_1}\,\varepsilon^{1/n_1} + o(\varepsilon^{1/n_1}),$$

where $\xi_1^1 = y_1^1 B x_1^1$. These two cases coincide when λ is simple.

THEOREM 2.2 (due to Lidskii [10]). *Let Φ_s be nonsingular for $s=1,\dots,q$ and let $j\in\{1,\dots,q\}$ be such that the r_j roots of (2.5) are different. Then the corresponding eigenvalues (2.6) of $A+\varepsilon B$ are simple for ε small enough and the associated right eigenvectors admit a power series expansion*

$$v_j^{kl}(\varepsilon) = u_j^k + \sum_{s=1}^{\infty} w_{js}^{kl}\,\varepsilon^{s/n_j}, \tag{2.7}$$

$k=1,\dots,r_j$, $l=1,\dots,n_j$, *where*

$$u_j^k = \sum_{p=1}^{f_j} c_j^{kp}\,x_j^p, \qquad k=1,\dots,r_j$$

and the column vector

$$c_j^k = \begin{bmatrix} c_j^{k1} \\ \vdots \\ c_j^{kf_j} \end{bmatrix}$$

satisfies

$$(\Phi_j - \xi_j^k E_j)\, c_j^k = 0. \tag{2.8}$$

Proof of Theorem 2.1. We may suppose, for the sake of simplicity, that A has only one eigenvalue λ; i.e., $\widehat{J}$ is empty. The general case may be reduced to this one using appropriate Riesz projections (we refer to Lidskii's original paper [10, pp. 83–84] or [1, section 3.9.1] for the details). Thus, we are interested in the roots ω of the characteristic equation

$$\det C(\omega, \varepsilon) \equiv \det(\omega I - J - \varepsilon \widetilde{B}) = 0, \qquad \widetilde{B} = P^{-1}BP. \tag{2.9}$$

As announced in section 1, we perform on $C(\omega, \varepsilon)$ the change of variables

$$z = \varepsilon^{1/n_j},$$

$$\mu = \frac{\omega - \lambda}{z},$$

where n_j is the partial multiplicity corresponding to j. This leads to a polynomial equation

$$\det \mathcal{P}(\mu, z) = \det[(\lambda + \mu z)I - J - z^{n_j}\widetilde{B}] = 0$$

in the new variables, where $\mathcal{P}(\mu, z) = C(\lambda + \mu z, z^{n_j})$. Since we are mainly concerned with solutions which are close to $z = 0$, it will prove convenient to multiply $\mathcal{P}(\mu, z)$ by the following diagonal matrices $L(z)$ and $R(z)$, partitioned conformally with J:

$$L(z) = \operatorname{Diag}[L_1^1, \ldots, L_1^{r_1}, \ldots, L_q^1, \ldots, L_q^{r_q}],$$
$$R(z) = \operatorname{Diag}[R_1^1, \ldots, R_1^{r_1}, \ldots, R_q^1, \ldots, R_q^{r_q}],$$

where

$$L_i^1(z) = \cdots = L_i^{r_i}(z) = \operatorname{diag}[z^{-1}, z^{-2}, \ldots, z^{-n_i}] \qquad \text{if } i \geq j,$$

$$L_i^1(z) = \cdots = L_i^{r_i}(z) = \operatorname{diag}[\underbrace{1, \ldots, 1}_{n_i - n_j}, z^{-1}, z^{-2}, \ldots, z^{-n_j}] \qquad \text{if } i < j$$

and

$$R_i^1(z) = \cdots = R_i^{r_i}(z) = \operatorname{diag}[1, z, z^2, \ldots, z^{n_i - 1}] \qquad \text{if } i \geq j,$$

$$R_i^1(z) = \cdots = R_i^{r_i}(z) = \operatorname{diag}[\underbrace{1, \ldots, 1}_{n_i - n_j}, 1, z, z^2, \ldots, z^{n_j - 1}] \qquad \text{if } i < j$$

for $i = 1, \ldots, q$ (note that $n_i \geq n_j$ if and only if $i \leq j$). We now introduce the matrix $F(\mu, z) = L(z)\,\mathcal{P}(\mu, z)\,R(z)$ and define

$$\mathcal{Q}(\mu, z) = \det\, F(\mu, z).$$

The nonsingularity of both $L(z)$ and $R(z)$ implies that, for any given $z \neq 0$,

$$\det\, \mathcal{P}(\mu, z) = 0 \;\Leftrightarrow\; \mathcal{Q}(\mu, z) = 0,$$

although of course the condition numbers of $L(z)$, $R(z)$ diverge to ∞ as $z \to 0$.

Let us show that $\mathcal{Q}$ is a polynomial in μ and z. For this, we split $F(\mu, z) = G(\mu, z) + H(z)$, where

$$G(\mu, z) = L(z)\,[(\lambda + \mu z)\,I - J]\,R(z)$$

is block diagonal and

$$H(z) = -z^{n_j}\,L(z)\,\widetilde{B}\,R(z).$$

We write $\Gamma_s^k = \lambda\,I + N_s$, $k = 1, \ldots, r_s$, where

$$N_s = \begin{bmatrix} 0 & 1 & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & 1 \\ & & & & 0 \end{bmatrix}$$

for $s = 1, \ldots, q$, and use two straightforward properties of the matrices L_i^k and R_i^k, namely, that

$$L_i^k(z)\,N_i\,R_i^k(z) = N_i, \qquad i = 1, \ldots, q, \quad k = 1, \ldots, r_i,$$
$$L_i^k(z)\,R_i^k(z) = z^{-1}\,I \qquad \text{whenever } n_i \leq n_j,$$

to check that the diagonal blocks $G_i^k(\mu, z) = L_i^k\,(\mu z\,I - N_i)\,R_i^k$ of $G(\mu, z)$ are

$$G_i^k(\mu, z) = \begin{cases} \mu\,I - N_i & \text{if } i \geq j, \\ \operatorname{diag}\,[\underbrace{\mu z, \ldots, \mu z}_{n_i - n_j}, \mu, \ldots, \mu] - N_i & \text{if } i < j. \end{cases} \tag{2.10}$$

Hence, all powers of z in G are nonnegative, and the same applies to H since no negative powers appear in either $z^{n_j}\,L_i^k$ or R_i^k. This proves our claim that $\mathcal{Q}(\mu, z)$ is a polynomial.

Let us now examine $F(\mu, 0) = G(\mu, 0) + H(0)$. The block diagonal matrix $G(\mu, 0)$ is given by equations (2.10) with $z = 0$. To give a careful description of $H(0)$ we need to partition $\widetilde{B}$, F, and H conformally with J. We denote by

$$\widetilde{B}\,{}_{j_1 j_2}^{k_1 k_2}, \qquad j_i = 1, \ldots, q, \quad k_i = 1, \ldots, r_{j_i}, \quad i = 1, 2$$

the $n_{j_1} \times n_{j_2}$ block of $\widetilde{B}$ lying on the same rows as $\Gamma_{j_1}^{k_1}$ and on the same columns as $\Gamma_{j_2}^{k_2}$ (the corresponding blocks of H and F are defined likewise). Following this notation, we have

$$H_{j_1 j_2}^{k_1 k_2}(z) = -z^{n_j} L_{j_1}^{k_1}(z)\, \widetilde{B}_{j_1 j_2}^{k_1 k_2}\, R_{j_2}^{k_2}(z),$$

which implies, in the first place, that

$$H_{j_1 j_2}^{k_1 k_2}(0) = 0 \quad \text{if } n_{j_1} < n_j \tag{2.11}$$

due to the vanishing of $z^{n_j} L_{j_1}^{k_1}$ at $z = 0$. On the other hand, if $j_1 \le j$ and $j_2 \le j$, an elementary calculation shows that

$$H_{j_1 j_2}^{k_1 k_2}(0) = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ -\beta_{j_1 j_2}^{k_1 k_2} & * & \cdots & * & 0 & \cdots & 0 \end{bmatrix}, \tag{2.12}$$

where the n_{j_2}–n_j elements marked with an asterisk are irrelevant to our argument and $\beta_{j_1 j_2}^{k_1 k_2}$ is the element in the lower left corner of the block $\widetilde{B}_{j_1 j_2}^{k_1 k_2}$. The same structure (2.12) applies to the case $j_1 \le j$, $j_2 > j$, but with zeros instead of asterisks. The main point of (2.12) is that every $\beta_{j_1 j_2}^{k_1 k_2} = y_{j_1}^{k_1} B\, x_{j_2}^{k_2}$ is an element of $\Phi_s = Y_s B X_s$ for $s = \max\{j_1, j_2\}$. In other words, we may find all the elements of the matrix Φ_j by looking at the lower left corners of the blocks (2.12) for $j_1, j_2 \le j$ (or, equivalently, for $n_{j_1}, n_{j_2} \ge n_j$).

Before our final examination of $\mathcal{Q}(\mu, 0)$, let us briefly turn to the diagonal blocks of $F(\mu, 0)$ of size n_j, i.e.,

$$F_{jj}^{kk}(\mu, 0) = \begin{bmatrix} \mu & -1 & 0 & \cdots & 0 \\ 0 & \mu & -1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ -\beta_{jj}^{kk} & 0 & 0 & \cdots & \mu \end{bmatrix}.$$

The determinant of this block does not change if we add to its first column the products of its second column by μ, of its third column by $\mu^2, \ldots$, and of its last column by $\mu^{n_j - 1}$. Neither does the whole determinant $\mathcal{Q}(\mu, 0)$ if we perform identical operations on the same columns of the whole matrix $F(\mu, 0)$ since, according to (2.11) and (2.12), the $n_j - 1$ columns of F which are being multiplied by powers of μ have no nonzero elements outside F_{jj}^{kk}. This amounts to replacing every block F_{jj}^{kk} with the block

$$\begin{bmatrix} 0 & -1 & 0 & \cdots & 0 \\ 0 & \mu & -1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ -\beta_{jj}^{kk} + \mu^{n_j} & 0 & 0 & \cdots & \mu \end{bmatrix}.$$

Let us now prove that the determinant $\mathcal{Q}(\mu, 0)$ of the matrix we finally obtain can be written in the form

$$\mathcal{Q}(\mu,0) = \pm\,\mu^{\alpha}\,\det\left(\Phi_j - \mu^{n_j}\,E_j\right) \tag{2.13}$$

for a suitable $\alpha \geq 0$. Although elementary, the proof is quite messy in the general case, so we will instead illustrate the strategy on a specific example. Take, for instance, the case $q = 3,\ j = 2,\ n_1 = 4,\ n_2 = 3,\ n_3 = 2,\ r_1 = 1,\ r_2 = 2,\ r_3 = 1$; i.e., $\mathcal{Q}(\mu,0)$ is the determinant of the 12×12 matrix

$$\left[\begin{array}{cccc|ccc|ccc|cc}
0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \mu & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \mu & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-\beta_{11}^{11} & * & 0 & \mu & -\beta_{12}^{11} & * & 0 & -\beta_{12}^{12} & 0 & 0 & -\beta_{13}^{11} & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \mu & -1 & 0 & 0 & 0 & 0 & 0 \\
-\beta_{21}^{11} & 0 & 0 & \mu & -\beta_{22}^{11} + \mu^3 & 0 & \mu & -\beta_{22}^{12} & 0 & 0 & -\beta_{23}^{11} & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu & -1 & 0 & 0 \\
-\beta_{21}^{21} & 0 & 0 & \mu & -\beta_{22}^{21} & 0 & 0 & -\beta_{22}^{22} + \mu^3 & 0 & \mu & -\beta_{23}^{21} & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu
\end{array}\right].$$

There are four rows in this matrix (first, fifth, eighth, and twelfth) containing one single nonzero element. In calculating the determinant of the matrix we may, therefore, remove the rows and columns corresponding to these four elements. This leaves an 8×8 matrix $M_1(\mu)$ such that $\mathcal{Q}(\mu,0) = \mu\,\det M_1(\mu)$. Because of the previous deletions, there are again four rows in M_1 with one single nonzero element. Eliminating the appropriate rows and columns of M_1, we get a 4×4 matrix M_2 and, finally, deleting one row and one column of M_2, a 3×3 matrix

$$M_3 = \left[\begin{array}{ccc}
-\beta_{11}^{11} & -\beta_{12}^{11} & -\beta_{12}^{12} \\
-\beta_{21}^{11} & -\beta_{22}^{11} + \mu^3 & -\beta_{22}^{12} \\
-\beta_{21}^{21} & -\beta_{22}^{21} & -\beta_{22}^{22} + \mu^3
\end{array}\right]$$

such that $\mathcal{Q}(\mu,0) = \mu^2\,\det M_3(\mu)$. But we know from (2.12) that the $\beta_{j_1 j_2}^{k_1 k_2}$ are just the elements of Φ_2. Hence, $M_3(\mu) = -\Phi_2 + \mu^3\,E_2$ and we obtain (2.13) with $\alpha = 2$. The same procedure goes through to the general case, exploiting in much the same way our knowledge of the block structure of the modified matrix $F(\mu,0)$.

Once we have $\mathcal{Q}(\mu,0)$ factorized as in (2.13), we note that its second factor $\det\left(\Phi_j - \mu^{n_j}\,E_j\right)$ is a polynomial of degree r_j in μ^{n_j}. This is trivial if $j = 1$ and a consequence of the nonsingularity of Φ_{j-1} if $j > 1$. We now take $\mathcal{Q}(\mu,z)$ as a polynomial in μ whose coefficients are continuous z-dependent functions. The continuous dependence of the roots of $\mathcal{Q}$ upon its coefficients guarantees the existence of exactly $r_j n_j$ continuous functions

$$\mu_j^{kl}(z) = (\xi_j^k)^{1/n_j} + o(1), \qquad k = 1,\ldots,r_j, \quad l = 1,\ldots,n_j,$$

describing all solutions of $\mathcal{Q}(\mu(z),z) = 0$ for z small enough (recall that some roots ξ_j^k of (2.5) might be zero if Φ_j is singular). Expansion (2.4) is obtained by returning to the original variables (λ,ε).

Finally, if all r_j roots ξ_j^k of equation (2.5) are known to be distinct, we may apply the implicit function theorem to (2.13) to show that the $\mu_j^{kl}(z)$ are in fact analytic functions of z, thus giving rise to the power series (2.6). □

Proof of Theorem 2.2. In the conditions of Theorem 2.2 it is clear that, for ε small enough, no eigenvalue (2.6) corresponding to j can possibly coincide with any of the eigenvalues (2.4) corresponding to Jordan blocks of different dimensions.

Furthermore, given one of these simple eigenvalues $\lambda_j^{kl}(\varepsilon)$, a right eigenvector associated with it may be constructed by taking as its components the m cofactors of the elements of a row of $A+\varepsilon B-\lambda_j^{kl}I$. This implies the analyticity of the eigenvector since the elements of this latter matrix are analytic functions of ε and the cofactors are simply sums of products of these elements (we recall that the eigenvector is unique up to constant multiples due to the simplicity of the eigenvalue).

Finally, let $e_j^{kl}(z)$ be a vector in the null space of $F_j^{kl}(z) = F(\mu_j^{kl}(z), z)$, where $\mu_j^{kl}(z) = (\lambda_j^{kl}(z) - \lambda)/z$. Dropping for simplicity both sub- and superscripts, we have

$$F(z)e(z) = L(z)\mathcal{P}(\mu(z), z)R(z)e(z) = 0,$$

which, due to the nonsingularity of $L(z)$ for $z \neq 0$, shows that $R(z)e(z)$ is a right eigenvector of $J + z^{n_j}\widetilde{B}$ associated with λ_j^{kl}. Hence, the first term $R(0)e(0)$ of its ε-expansion must be, up to a constant, equal to the zero-th order term u_j^k of expansion (2.7). Equation (2.8) is finally obtained by applying to the linear system $F(0)e = 0$ the same ideas we used to simplify $\mathcal{Q}(\mu, 0)$ in the proof of Theorem 2.1. □

Remarks.

1. A proof of the existence of power series expansions (2.6) and (2.7) under the conditions of Theorem 2.2 goes back to Vishik and Lyusternik [13, Theorem 6, Appendix I]. Their approach, however, is radically different from Lidskii's, since they *impose* both expansions (2.6) and (2.7) as formal series at the outset, recursively find all coefficients, and finally prove the convergence of the series on some nontrivial interval around $\varepsilon = 0$. In their setting, the assumption that all Schur complements have nonzero distinct eigenvalues arises as a solvability condition on the system of infinitely many equations determining the coefficients of the series. Lidskii's approach in [10], on the other hand, concentrates only on the leading term, regardless of the rest of the expansion. This allows him to get more general results, avoiding at the same time the issue of convergence of the series: in those cases in which a power series expansion is obtained, its convergence is a consequence of the function theoretical results invoked in the proof.

2. The computation of Schur complements is equivalent to (and may be replaced by) choosing the eigenvector matrices X_j and Y_j in a special way. Suppose, for instance, that $j = 2$ and $\Phi_1 = Y_1 B X_1$ is nonsingular. A straightforward calculation shows that the columns of

$$\widetilde{X}_2 = X_2 - X_1\Phi_1^{-1}Y_1 B X_2 \tag{2.14}$$

and the rows of

$$\widetilde{Y}_2 = Y_2 - Y_2 B X_1 \Phi_1^{-1} Y_1 \tag{2.15}$$

are, respectively, right and left eigenvectors of A, corresponding to Jordan chains of the same length as the eigenvectors given by the rows and columns of X_2 and Y_2.

Furthermore, we have $Y_1 B \widetilde{X}_2 = 0$ and $\widetilde{Y}_2 B X_1 = 0$. Hence, it suffices to define

$$\widetilde{W}_2 = \begin{bmatrix} Y_1 \\ \widetilde{Y}_2 \end{bmatrix}, \qquad \widetilde{Z}_2 = \begin{bmatrix} X_1, \widetilde{X}_2 \end{bmatrix}, \quad \text{and} \quad \widetilde{\Phi}_2 = \widetilde{W}_2 B \widetilde{Z}_2$$

to obtain a block diagonal matrix $\widetilde{\Phi}_2$ whose lower right block $\widetilde{Y}_2 B \widetilde{X}_2$ is precisely the Schur complement of Φ_1 in the old matrix Φ_2. The replacement of X_3 and Y_3 by suitable matrices leads to the block diagonalization of Φ_3, provided that $\widetilde{\Phi}_2$ is nonsingular. It should be noted that, although only one of the two matrices (2.14) or (2.15) is required to reproduce the Schur complement for $j = 2$ (by block triangularizing Φ_2), both of them are needed to continue to the following step $j = 3$.

3. Both matrices Φ_{j-1} and Φ_j must be nonsingular to obtain the leading nonzero terms in all expansions (2.4). That is probably why Lidskii's original statement of Theorem 2.1 imposed nonsingularity of *both* Φ_{j-1} and Φ_j. Nevertheless, as we have seen in the proof of the theorem, only Φ_{j-1} need be nonsingular for the eigenvalue expansions (2.4) to hold: suppose that Φ_{j-1} is nonsingular and Φ_j is singular. Then we have

$$\det(\Phi_j - \xi E_j) = \xi^\beta \, q(\xi), \qquad q(0) \neq 0,$$

for a certain $\beta > 0$; i.e., (2.5) has β zero and $r_j - \beta$ nonzero solutions. Hence, $(r_j - \beta) n_j$ expansions (2.4) have a nonzero first-order term, while all we can say about the remaining βn_j eigenvalues is that they are of the form $\lambda_j^k(\varepsilon) = \lambda + o(\varepsilon^{1/n_j})$. This strongly suggests the possibility of interaction with eigenvalues associated with Jordan blocks of size less than n_j. These interactions will become much clearer in the next section with the use of the Newton diagram.

4. An important advantage of Lidskii's proof technique is that it does not require the analyticity of the perturbation. Consequently, Lidskii's approach can be used to investigate the variational behavior of eigenvalues under very weak differentiability hypotheses. For example, Theorem 2.1 remains valid for perturbations of class C^1. More generally, one can even obtain *one-sided* or *directional* versions of Theorem 2.1. For example, if $A : \mathbb{R} \to \mathbb{C}^{n \times n}$ is continuous at the origin with

$$A(\varepsilon) = A + \varepsilon B + o(\varepsilon) \quad \text{for } \varepsilon > 0,$$

then the expansion (2.4) in Theorem 2.1 remains valid for $\varepsilon \geq 0$. In fact, the same proof holds with minor changes. First observe that the continuity of the Riesz projections depends only on the continuity of the perturbation [7, Theorem 5.1]. Next, using a bar to denote matrices arising in this new setting, we find that

$$\det \overline{C}(\omega, \varepsilon) \equiv \det(\omega I - A - \varepsilon B - o(\varepsilon)) = \det C(\omega, \varepsilon) + o(\varepsilon) \quad \text{for } \varepsilon \geq 0,$$

where $C(\omega, \varepsilon)$ is given by (2.9). After changing to variables μ, z, we have

$$\det \overline{\mathcal{P}}(\mu, z) = \det \mathcal{P}(\mu, z) + o(z^{n_j}),$$

so, multiplying by $L(z)$ and $R(z)$, we obtain

$$\overline{F}(\mu, z) = L(z) \overline{\mathcal{P}}(\mu, z) R(z) = F(\mu, z) + \overline{S}(z)$$

for $\overline{S}(z) = L(z) \, M \, R(z)$, where M is $o(z^{n_j})$. Now, recall that no negative power of z in $L(z)$ has absolute value larger than n_j. This means that $\overline{S}(z) = o(1)$, so

$$\overline{\mathcal{Q}}(\mu, 0) = \det \overline{F}(\mu, 0) = \mathcal{Q}(\mu, 0)$$

and the factorization (2.13) still holds. Finally, although in this case $\overline{\mathcal{Q}}(\mu, z)$ is no longer a polynomial in both variables μ and z, it is a polynomial in μ, whose coefficients are continuous functions of z. Thus, we may still guarantee that the roots of $\overline{\mathcal{Q}}(\mu, z)$ depend continuously on z to conclude the proof.

Example. We consider the simplest case of a matrix having an eigenvalue which is neither semisimple nor nonderogatory: let A be a 3×3 matrix with a triple eigenvalue λ of geometric multiplicity two (in our notation, $q = 2$, $n_1 = 2$, $n_2 = 1$, $r_1 = r_2 = 1$). Dropping the superscripts, we denote the two left eigenvectors by y_1, y_2 and the two right eigenvectors by x_1, x_2. We find that $\Phi_1 = \beta_{11} = y_1 B x_1$ and

$$\Phi_2 = \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix}, \quad \beta_{ij} = y_i B x_j, \quad i, j = 1, 2.$$

We have two eigenvalues

$$\lambda_1^l(\varepsilon) = \lambda \pm \sqrt{\beta_{11}}\, \varepsilon^{1/2} + o(\varepsilon^{1/2}), \qquad l = 1, 2.$$

Furthermore, if both Φ_1 and $\det \Phi_2$ are different from zero, the third eigenvalue is $\lambda + \xi_2 \varepsilon + o(\varepsilon)$, where $\xi_2 = (\det \Phi_2)/\Phi_1$ is the solution of

$$\det(\Phi_2 - \xi E_2) = \det \Phi_2 - \Phi_1 \xi = 0.$$

Note that, if Φ_1 is zero, we know only that two of the eigenvalues are $o(\varepsilon^{1/2})$ perturbations of λ, without any further indication of their asymptotic order.

Note that even in this simple example it is unclear which leading powers of ε are to be expected when some Φ_j is singular. Lidskii [10] gives an example where all three eigenvalues above are perturbed by order $\varepsilon^{2/3}$, and we have seen in (1.1) a similar example of a 5×5 matrix A with Jordan blocks of sizes 3 and 2, whose perturbed eigenvalues are of order $\varepsilon^{2/5}$ for a conveniently chosen perturbation. None of these leading exponents can be explained, in principle, by any of the above results. It seems that the information that $\mathcal{Q}(\mu, 0)$ provides in the case of singular Φ_j is helpful only in predicting which powers of ε *cannot* appear in the eigenvalue expansions. In the following section we present an alternative approach that will improve our understanding of Theorem 2.1, giving us a global picture of what happens in the degenerate case when some Φ_j is singular.

3. Application of Newton's diagram. In this section the symbol λ is used as a parameter, not as a fixed value as earlier. We consider a complex polynomial equation

$$P(\lambda, \varepsilon) = \lambda^m + \alpha_1(\varepsilon)\lambda^{m-1} + \cdots + \alpha_{m-1}(\varepsilon)\lambda + \alpha_m(\varepsilon) = 0 \tag{3.1}$$

in λ, with analytic coefficients

$$\alpha_k(\varepsilon) = \hat{\alpha}_k \varepsilon^{a_k} + \cdots, \qquad k = 1, \ldots, m,$$

where a_k is the leading exponent and $\hat{\alpha}_k$ the leading coefficient of $\alpha_k(\varepsilon)$ (i.e., $\hat{\alpha}_k \neq 0$ and no term of order lower than a_k appears in the expansion of $\alpha_k(\varepsilon)$). It is well

known [1], [7] that the roots of (3.1) are given by expansions in fractional powers of ε. The leading exponents of these expansions can be easily found through the following elementary geometrical construction: we plot the values a_k versus k for $k = 1, \ldots, m$ together with the point $(0, 0)$ corresponding to λ^m (if $\alpha_j(\varepsilon) \equiv 0$, the corresponding point is disregarded). Then we draw the segments on the lower boundary of the convex hull of the plotted points. These segments constitute the so-called *Newton diagram* associated with $P(\lambda, \varepsilon)$ (see Fig. 3.1 for two specific examples). One can prove [1, Appendix A7], [2], [12] that the slopes of the different segments of the Newton diagram are precisely the leading powers of the ε-expansions of the roots $\lambda = \lambda(\varepsilon)$ of (3.1). The number of roots corresponding to each slope equals the length of the projection on the horizontal axis of the segment with that particular slope. The underlying idea is to substitute an *Ansatz*

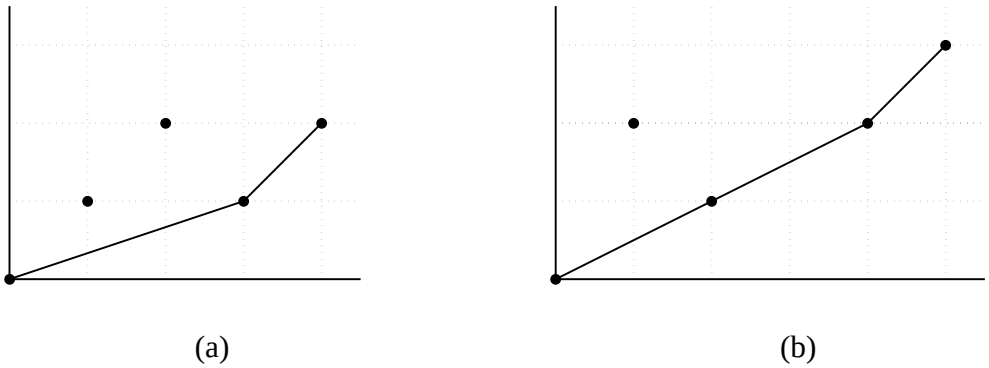


FIG. 3.1. *Newton diagrams associated with the polynomials:* (a) $\lambda^4 + (2\varepsilon - \varepsilon^2)\lambda^3 + \varepsilon^2\lambda^2 + (\varepsilon - \varepsilon^3)\lambda + \varepsilon^2$, (b) $\lambda^5 - \varepsilon^2\lambda^4 + (\varepsilon - 3\varepsilon^2)\lambda^3 + \varepsilon^2\lambda - \varepsilon^3$.

$$\lambda(\varepsilon) = \mu\varepsilon^{\beta} + \cdots \tag{3.2}$$

into (3.1), with μ, β to be determined. Every point (k, a_k) plotted in the diagram produces an $\varepsilon^{a_k+(m-k)\beta}$ term. If $\lambda(\varepsilon)$ is a root of (3.1), all the terms we obtain from this substitution must cancel each other. Hence, at least two terms of the lowest order in ε must be present, and this lowest order is clearly to be found among the exponents $\{a_1 + (m-1)\beta, a_2 + (m-2)\beta, \ldots, a_m\}$. Consider the segment S of the Newton diagram with the smallest slope s and choose $\beta = s$ in (3.2). All points (k, a_k) lying on S give rise to terms with the same exponent since $a_k + (m-k)s$ is constant on S. The fact that no point (k, a_k) lies below S implies that no other term of the expansion can be of a lower order in ε. Hence, the leading coefficients μ are determined as the solutions of

$$\sum_{(k,a_k)\in S} \mu^{m-k}\hat{\alpha}_k = 0. \tag{3.3}$$

We get the leading terms of the remaining roots of (3.1) by repeating the same argument for the increasingly larger slopes appearing in the Newton diagram.

Returning to the eigenvalue problem, our main goal in this section is to establish the relationship between the quantities $\det \Phi_j$, $j = 1, \ldots, q$, and the Newton diagram associated with the characteristic polynomial of $A + \varepsilon B$. We recall that A is a matrix

with only one eigenvalue (previously denoted by λ) of multiplicity m, with partial multiplicities n_j, each repeated r_j times for $j = 1, \ldots, q$. With no loss of generality, we may assume that this eigenvalue is zero. In this case, the characteristic polynomial $p(\lambda, \varepsilon) = \det(\lambda I - A - \varepsilon B) = \det(\lambda I - J - \varepsilon \tilde{B})$ can be written in the form (3.1), with $m = \sum_1^q r_j n_j$ (recall that λ is no longer an eigenvalue of A, but the unknown in the characteristic polynomial). To draw the Newton diagram associated with $p(\lambda, \varepsilon)$, we must know the exponents a_k for $k = 1, \ldots, m$. This is quite easy if the eigenvalue is semisimple, since the Jordan form J of A is zero and each $\alpha_k(\varepsilon)$ equals ε^{m-k} multiplied by a certain sum of minors of $\tilde{B}$ of dimension $m - k$. In this case, the Newton diagram is formed by one single segment of slope $s = 1$. If the eigenvalue is not semisimple, some nontrivial Jordan block appears in J, which means that, apart from the $O(\varepsilon^{m-k})$ terms, each $\alpha_k(\varepsilon)$ will typically contain terms of lower order generated by the ones appearing above the diagonal of J. This clearly shows that the effect of nontrivial Jordan blocks is to introduce in the Newton diagram line segments with slopes less than 1, with the smallest possible slope corresponding to the case of a nonderogatory eigenvalue (one single segment of slope $1/m$) and the largest possible one to the semisimple case. All possible Newton diagrams for the given multiplicity m lie between these two extremal segments.

We must now carefully determine which points (k, a_k) may appear on the Newton diagram for a particular Jordan structure. To this purpose, it will be useful to find the lowest possible diagram compatible with the given Jordan structure. To do this, we fix every exponent l of ε and find the largest possible $k = k(l)$ such that there exists a perturbation matrix B for which $a_{k(l)} = l$. This amounts to fixing a height l on the vertical axis of the Newton diagram and determining the rightmost possible point $(k(l), l)$ in the diagram. The following theorem gives us the values $k(l)$ for the exponents $l = 1, \ldots, f_q$ which are relevant to our argument (we recall that $f_j = r_1 + \cdots + r_j$) and, more importantly, also provides some coefficients of the characteristic polynomial which are crucial to determine the Newton diagram.

THEOREM 3.1. *For every l, $l = 1, \ldots, f_q$, the corresponding $k(l)$ is equal to the sum of the dimensions of the l largest Jordan blocks of J. More precisely, write $f_0 = 0$ and suppose $l = f_{j-1} + \rho$ for some $j = 1, \ldots, q$ and $0 < \rho \le r_j$. Then*

$$k(l) = r_1 n_1 + \cdots + r_{j-1} n_{j-1} + \rho\, n_j$$

and the coefficient of ε^l in $\alpha_{k(l)}$ is equal to $(-1)^l$ multiplied by the sum of all principal minors of Φ_j corresponding to submatrices of dimension l that contain the upper left block Φ_{j-1} of Φ_j (if $j = 1$, all principal minors of dimension l are to be considered). If, in particular, $l = f_j$ for some $j \in \{1, \ldots, q\}$, then the coefficient of ε^{f_j} in $\alpha_{k(f_j)}$ is $(-1)^{f_j} \det \Phi_j$.

Proof. The characteristic polynomial of $A + \varepsilon B$ is a linear combination with coefficients ± 1 of all possible products of m elements of the matrix $\lambda I - J - \varepsilon \widetilde{B}$, with the restriction that no two factors can be on the same row or the same column.

It is clear that the only way to obtain a product of order ε^l is to choose exactly $m - l$ factors containing ε-independent terms (i.e., "lambdas" or "minus ones"). Furthermore, we should try to include as few lambdas as possible among these factors, since we are looking for the smallest possible power of λ. However, we are not free to make whatever choices we want. Due to the special position of the ε-independent terms, every time we choose a minus one we are, at the same time, excluding from the product those lambdas which lie on the same row or the same column. Let us

examine the restrictions. Suppose an admissible choice (i.e., a choice producing a term of order ε^l) contains β minus ones. These β choices remove a certain number of lambdas, which depends on the number of Jordan blocks these β minus ones are sampled from. This is due to the fact that the *first* minus one we choose from a particular block excludes two lambdas, while any further minus one from the same block removes only one. Suppose the β minus ones were taken from γ different blocks. These β choices exclude $\beta+\gamma$ lambdas, which, together with the $m-l-\beta$ lambdas which were actually chosen in the product, cannot exceed the total number m of available lambdas. We conclude that $\gamma \leq l$; i.e., we are allowed to sample minus ones from at most l Jordan blocks. Hence, to produce the lowest possible power of λ we must exhaust all minus ones from the l largest possible Jordan blocks of J, and only then complete with lambdas until we have the $m-l$ factors.

Suppose, in the first place, that $l = f_j$ for some $j = 1, \ldots, q$. Then there is only one way of choosing these $m-l$ factors: we must choose the $\beta = r_1(n_1-1) + \cdots + r_j(n_j-1)$ minus ones from the l largest Jordan blocks, plus the $m-l-\beta = r_{j+1}n_{j+1} + \cdots + r_q n_q$ diagonal lambdas from the remaining Jordan blocks. Thus, we get $k(l) = l+\beta = r_1 n_1 + \cdots + r_j n_j$. Note that if we delete from $\lambda I - J - \varepsilon\tilde{B}$ the rows and columns corresponding to these $m-l$ elements, the $f_j \times f_j$ remaining matrix is precisely $-\varepsilon\Phi_j$, which proves our claim.

The same argument is valid in the case $\rho < r_{j+1}$, although in this case there is more than one way of building up the products: each one corresponds to a different choice of ρ blocks among the r_{j+1} Jordan blocks of dimension n_{j+1}, generating a different principal minor of Φ_j to be included in the sum. □

Let us now introduce the following definition.

DEFINITION 3.2. *Denote* $\mathcal{P}_j \equiv (k(f_j), f_j)$, *and let* S_j *be the segment of slope* $1/n_j$ *connecting* $\mathcal{P}_{j-1}$ *with* $\mathcal{P}_j$ *for* $j = 1, \ldots, q$. *We define the* Newton envelope corresponding to the Jordan structure of J *as the diagram obtained by successively joining the segments* $S_1, S_2, \ldots, S_q$.

As a first consequence of Theorem 3.1, note that all points $(k(l), l)$ for l between f_{j-1} and f_j lie along the corresponding segment S_j. Hence, the Newton envelope is indeed the lowest Newton diagram we were looking for. This is not, however, its most interesting feature. Keep in mind that, *given a particular* B, only those points $(k(l), l)$ from the envelope such that $a_{k(l)} = l$ will actually be plotted in the Newton diagram. This means, in particular, that a corner point $\mathcal{P}_j$ of the Newton envelope appears on the Newton diagram only if the perturbation B is such that the corresponding coefficient $\pm \det \Phi_j$ is nonzero. In other words, the Newton envelope displays the *generic* behavior of the eigenvalues of A under perturbation, in the sense that it coincides with the Newton diagram in all situations except in those nongeneric cases in which the perturbation B causes one of the Φ_j to be singular.

Theorem 3.1 largely explains the importance of the $\det \Phi_j$ in obtaining the exponents $1/n_j$ in the eigenvalue expansions. Furthermore, it clears the way for an independent proof of Lidskii's Theorem 2.1.

Proof of Theorem 2.1. Let us suppose first that $j \in \{2, \ldots, q\}$ is such that Φ_{j-1} and Φ_j are nonsingular (the case $j = 1$ is completely analogous). Then both $\mathcal{P}_{j-1}$ and $\mathcal{P}_j$ are among the points plotted to construct the Newton diagram and S_j is one of the segments in the diagram (no point (k, a_k) can lie below S_j). Thus, we obtain the leading exponent of expansion (2.4). We also get the leading coefficient by carefully examining, for the segment S_j, equation (3.3). We first note that the only candidates (k, a_k) to lie on S_j are the intermediate points $\mathcal{Q}_t \equiv (k(f_j - t), f_j - t)$, $t =$

$1, \ldots, r_j - 1$. The fact that the $\mathcal{Q}_t$ are separated from each other by a distance n_j on the horizontal axis implies that equation (3.3) depends on μ only through μ^{n_j}. More precisely, let T be the set of values $t \in \{1, \ldots, r_{j-1}\}$ such that $\mathcal{Q}_t$ appears in the Newton diagram. Then

$$\sum_{(k,a_k)\in S_j} \mu^{m-k}\hat{\alpha}_k = \mu^{m-k(f_j)} \left[\mu^{n_j r_j} \hat{\alpha}_{k(f_{j-1})} + \sum_{t\in T} \hat{\alpha}_{k(f_j - t)} \mu^{t\, n_j} + \hat{\alpha}_{k(f_j)} \right] = 0,$$

where the expression in brackets is a polynomial in μ^{n_j}. Now, we recall from Theorem 3.1 that for each $l = f_j - t$ with $t \in T$, the corresponding $\hat{\alpha}_{k(l)}$ is (up to the sign) the sum of all principal minors of Φ_j of dimension l containing Φ_{j-1}, which is precisely the way the coefficients of the powers of ξ are obtained in $\det(\Phi_j - \xi E_j)$. This implies that we get the nonzero solutions of (3.3) by solving $\det(\Phi_j - \mu^{n_j} E_j) = 0$.

Now suppose that Φ_j is singular. Then the corresponding point $\mathcal{P}_j$ no longer belongs to the diagram, implying the loss of some of the expansions (2.4) or, equivalently, the loss of part of the segment S_j. The part of S_j that actually remains depends upon the nullity of Φ_j. If rank $\Phi_j = f_j - \beta$, there are β rows or columns of Φ_j that depend linearly on the remaining ones. This means, on one hand, that no point $\mathcal{Q}_t \equiv (k(f_j - t), f_j - t)$ appears in the diagram for $1 \le t < \beta$ and, on the other hand, that $\mathcal{Q}_\beta$ does appear (each principal minor of Φ_j of dimension $f_j - \beta$ either vanishes or takes a common nonzero value, since only $f_j - \beta$ columns of Φ_j are linearly independent). We conclude that the part of S_j remaining in the Newton diagram is the segment connecting $\mathcal{P}_{j-1}$ to $\mathcal{Q}_\beta = (k(f_j - \beta), f_j - \beta)$ (see Fig. 3.2). This accounts for $(r_j - \beta)n_j$ expansions (2.4), whose leading coefficients are,

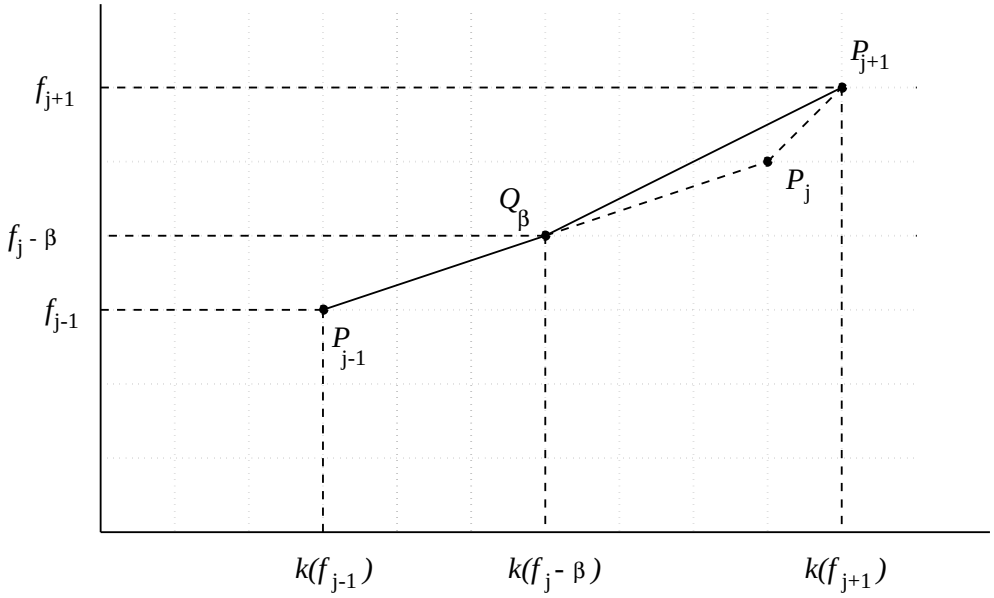


FIG. 3.2. *The Newton diagram is shown as a solid line and the envelope as a dashed line.*

reasoning as above, the n_jth roots of the $r_j - \beta$ nonzero solutions of equation (2.5). As for the βn_j remaining eigenvalues, they correspond to segments whose slope is strictly larger than $1/n_j$. Hence, the remaining expansions (2.4) are still valid since they correspond to the β null solutions of equation (2.5). □

A further consequence of this Newton diagram approach is that the hypotheses of Theorem 2.2 can be weakened in the sense that, if j is such that all roots of (2.5) are distinct, we only need $\Phi_1, \ldots, \Phi_j$ to be nonsingular to guarantee the simplicity of the $r_j n_j$ eigenvalues $\lambda_j^{kl}(\varepsilon)$: the slope of any segment of the diagram lying to the right of $\mathcal{P}_j$ is strictly larger than $1/n_j$ regardless of the singularity of Φ_s, $s = j+1, \ldots, q$. Hence, no eigenvalue corresponding to a Jordan block of size less than n_j can coincide with any $\lambda_j^{kl}(\varepsilon)$ if ε is small enough.

Apart from recovering the results of section 2, the approach through the Newton diagram is quite helpful in getting a better understanding of the nongeneric case when some Φ_j is singular. The fact that in this case $\mathcal{P}_j$ does not belong to the Newton diagram implies that neither the complete segment S_j nor the complete segment S_{j+1} appears on the Newton diagram. This indicates some kind of interaction between the eigenvalues associated with blocks of size n_j and those associated with blocks of size n_{j+1}. We may, in fact, use the Newton diagram as a tool to quantify this interaction, actually finding both leading exponents and coefficients of the missing eigenvalue expansions in simple situations. The range of possibilities is easily visualized with the aid of the Newton envelope. For example, if (h_1, k_1) and (h_2, k_2) are two points that are known to lie on both the Newton diagram and the Newton envelope, then the segment of the Newton diagram between h_1 and h_2 must necessarily lie between the chord connecting (h_1, k_1) to (h_2, k_2) and the Newton envelope. Thus, to determine the Newton diagram one need only focus on the integer lattice points trapped between this chord and the Newton envelope. As an illustration of the power of this observation in the nongeneric case, we provide the following corollary. In this corollary, we identify a case in which no integer lattice points can lie between the chord and the Newton envelope.

COROLLARY 3.3. *Let* $0 \le \beta \le r_j$ *and* $0 \le \alpha \le r_{j+1}$. *Suppose that* $\mathcal{Q}_\beta^j = (k(f_j - \beta), f_j - \beta)$ *and* $\hat{\mathcal{Q}}_\alpha^j = (k(f_j + \alpha), f_j + \alpha)$ *are two points lying on the Newton diagram, while the points* $\mathcal{Q}_s^j$ *for* $s = \beta - 1, \ldots, 1$, $\mathcal{P}_j$ *and* $\hat{\mathcal{Q}}_t^j$ *for* $t = 1, 2, \ldots, \alpha - 1$ *do not lie on the Newton diagram. Set* $p = \beta n_j + \alpha n_{j+1}$ *and* $\sigma = \frac{\beta+\alpha}{p}$. *If*

$$(\sigma n_j - 1)\beta \le \min(\sigma, 1 - \sigma), \tag{3.4}$$

then there are p *eigenvalues of* $A + \varepsilon B$ *of the form*

$$\lambda^l(\varepsilon) = \lambda + \eta^{1/p}\varepsilon^\sigma + o(\varepsilon^\sigma), \qquad l = 1, 2, \ldots, p, \tag{3.5}$$

where $\eta \neq 0$. *Moreover, if either*

(i) *the inequality in* (3.4) *is strict or*

(ii) $(\sigma n_j - 1)\beta < \sigma$ *and* $\alpha = n_{j+1} = 1$,

then

$$\eta = -\frac{\hat{\alpha}_{k(f_j+\alpha)}}{\hat{\alpha}_{k(f_j-\beta)}}\ . \tag{3.6}$$

Proof. As noted above, the Newton diagram must lie between the chord connecting $\mathcal{Q}_\beta^j$ to $\hat{\mathcal{Q}}_\alpha^j$ and the Newton envelope on the interval $[k(f_j - \beta), k(f_j + \alpha)]$. Furthermore, since none of the points $\mathcal{Q}_s^j$ for $s = \beta - 1, \ldots, 1$, $\mathcal{P}_j$ and $\hat{\mathcal{Q}}_t^j$ for $t = 1, 2, \ldots, \alpha - 1$ lie on the Newton diagram, the Newton envelope and diagram coincide only at the end points of this interval. Thus, the expansions (3.5) will be valid if we can show that there are no integer lattice points strictly between the chord and the Newton

diagram on the interval $[k(f_j - \beta), k(f_j + \alpha)]$. To do this we need only show that the lattice points $(k(f_j) - 1, f_j)$ and $(k(f_j) + 1, f_j + 1)$ lie on or above the chord. The condition that $(k(f_j) - 1, f_j)$ lies on or above the chord yields the inequality $(\sigma n_j - 1)\beta \leq \sigma$, while the condition that $(k(f_j)+1, f_j+1)$ lies on or above the chord yields the inequality $(\sigma n_j - 1)\beta \leq 1 - \sigma$. Note that this second condition is no longer needed if $\alpha = n_{j+1} = 1$, since in this case $(k(f_j)+1, f_j+1)$ coincides with $\hat{\mathcal{Q}}^j_\alpha$. Thus, under either condition (i) or (ii), $\mathcal{Q}^j_\beta$ and $\hat{\mathcal{Q}}^j_\alpha$ are the only integer lattice points on the chord and so (3.6) follows from (3.3). □

It is interesting to consider a few special cases of the above result. Note that if $\beta = 0$, then α can take any of the values $0, 1, \ldots, r_{j+1}$, and if $\alpha = 0$, then β can take any of the values $0, 1, \ldots, r_j$. The case $\beta = 0$ reaffirms the third remark at the end of section 2, while the case $\alpha = 0$ illustrates that expansions with power $1/n_j$ are possible even if Φ_{j-1} is singular. Finally, if one is given a fixed value for either α or β, then simple bounds on the other value are easily obtained from (3.4). For example, if $\alpha = 1$, then the restriction (3.4) yields the inequality $\beta(n_j - n_{j+1} - 1) \leq 1$. That is, if $n_j = n_{j+1} + 1$, then β can take on any of the values $0, 1, \ldots, r_j$; if $n_j = n_{j+1} + 2$, then β can take only the values 0 and 1; and if $n_j > n_{j+1} + 2$, then β must be zero.

Corollary 3.3 also explains the exponent appearing in the eigenvalues of Wilkinson's example (1.1). In that case, the point $\mathcal{P}_1 = (3, 1)$ does not lie on the Newton diagram since $\Phi_1 = 0$. On the other hand, $\mathcal{P}_2 = (5, 2)$ does appear, due to the nonsingularity of

$$\Phi_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Thus, an application of Corollary 3.3 with $j = 1$, $\alpha = \beta = 1$ shows that there are $p = 5$ eigenvalues of order $\sigma = 2/5$.

The situation becomes more complicated with the introduction of more integer lattice points between the chord and the Newton envelope, since more possibilities for the Newton diagram arise. But, in many cases, these possibilities can be delineated by considering certain key lattice points as was done in the proof of the above corollary. Indeed, this approach can provide a fairly complete picture in many particular cases. Let us conclude by applying the ideas of this section to some specific examples.

Example 1. We first turn to our example in the preceding section of a 3×3 matrix with $q = 2$, $n_1 = 2$, $n_2 = 1$, $r_1 = r_2 = 1$. The expansions we obtained in the previous section correspond to a Newton diagram with two segments: S_1 connecting $(0, 0)$ with $\mathcal{P}_1 = (2, 1)$ and S_2 joining $\mathcal{P}_1$ and $\mathcal{P}_2 = (3, 2)$. Note that this is precisely the Newton envelope corresponding to the given Jordan structure (see Fig. 3.3(a)). Now suppose that $\Phi_1 = 0$ with $\det \Phi_2 \neq 0$. This means that $\mathcal{P}_1$ no longer is plotted, so the diagram consists of a single segment of slope $2/3$ joining $(0, 0)$ with $\mathcal{P}_2$ (see Fig. 3.3(b)). If both Φ_1 and Φ_2 are singular, we obtain one single segment of slope 1 as long as B is nonsingular.

Example 2. We consider a 5×5 matrix with one zero eigenvalue and $q = 2$, $n_1 = 2$, $n_2 = 1$, $r_1 = 2$, $r_2 = 1$. We assume, for the sake of simplicity, that A is already in Jordan form, i.e.,

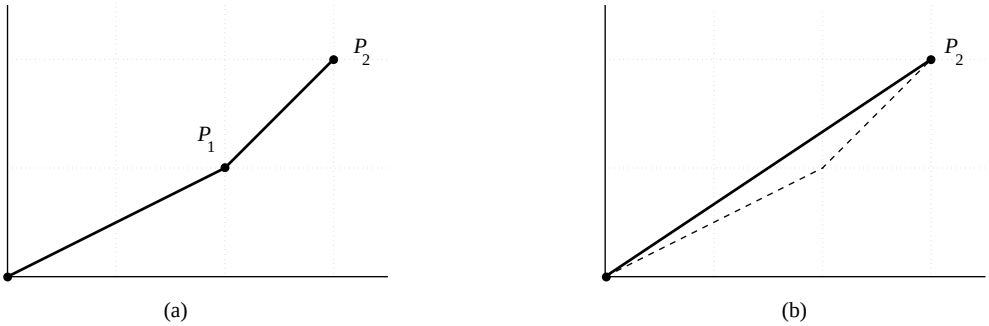


FIG. 3.3. *Newton diagrams corresponding to Example* 1. *In* (a) *the Newton diagram and envelope coincide. In* (b) *the diagram is shown as a solid line and the envelope as a dashed line.*

$$A = \left[\begin{array}{cc|cc|c} 0 & 1 & & & \\ & 0 & & & \\ \hline & & 0 & 1 & \\ & & & 0 & \\ \hline & & & & 0 \end{array}\right].$$

If $B = (b_{ij})_{i,j=1}^{5}$, then

$$\Phi_1 = \begin{bmatrix} b_{21} & b_{23} \\ b_{41} & b_{43} \end{bmatrix}, \qquad \Phi_2 = \left[\begin{array}{cc|c} b_{21} & b_{23} & b_{25} \\ b_{41} & b_{43} & b_{45} \\ \hline b_{51} & b_{53} & b_{55} \end{array}\right].$$

We consider the different possibilities.

(i) Suppose $\det \Phi_1 \neq 0$, so that $\mathcal{P}_1 = (4, 2)$ appears in the diagram. Then the perturbed matrix has four eigenvalues

$$\lambda_1^{kl}(\varepsilon) = (\xi_1^k)^{1/2}\varepsilon^{1/2} + o(\varepsilon^{1/2}), \qquad k, l = 1, 2,$$

where ξ_1^1, ξ_1^2 are the eigenvalues of Φ_1.

• If, additionally, $\det \Phi_2 \neq 0$, so that $\mathcal{P}_2 = (5, 3)$ also appears in the diagram, then the fifth eigenvalue of $A + \varepsilon B$ is $\lambda_2^{11}(\varepsilon) = \xi_2\varepsilon + o(\varepsilon)$ for

$$\xi_2 = b_{55} - \begin{bmatrix} b_{51} & b_{53} \end{bmatrix} \Phi_1^{-1} \begin{bmatrix} b_{25} \\ b_{45} \end{bmatrix}.$$

In this case, the Newton diagram and envelope coincide (see Fig. 3.4 (a)).

• If, on the other hand, Φ_2 is singular, $\mathcal{P}_2$ does not appear in the diagram. In this case, the order of the fifth eigenvalue is at least $O(\varepsilon^2)$, corresponding to a segment joining $\mathcal{P}_1 = (4, 2)$ with $(5, 4)$. However, higher slopes might appear in some cases.

(ii) Suppose now that Φ_1 is singular. Then the point $\mathcal{P}_1$ no longer appears in the Newton diagram, and to determine the order of the perturbations we need to know whether or not the coefficient

$$\delta = -\operatorname{tr} \Phi_1 = -b_{21} - b_{43}$$

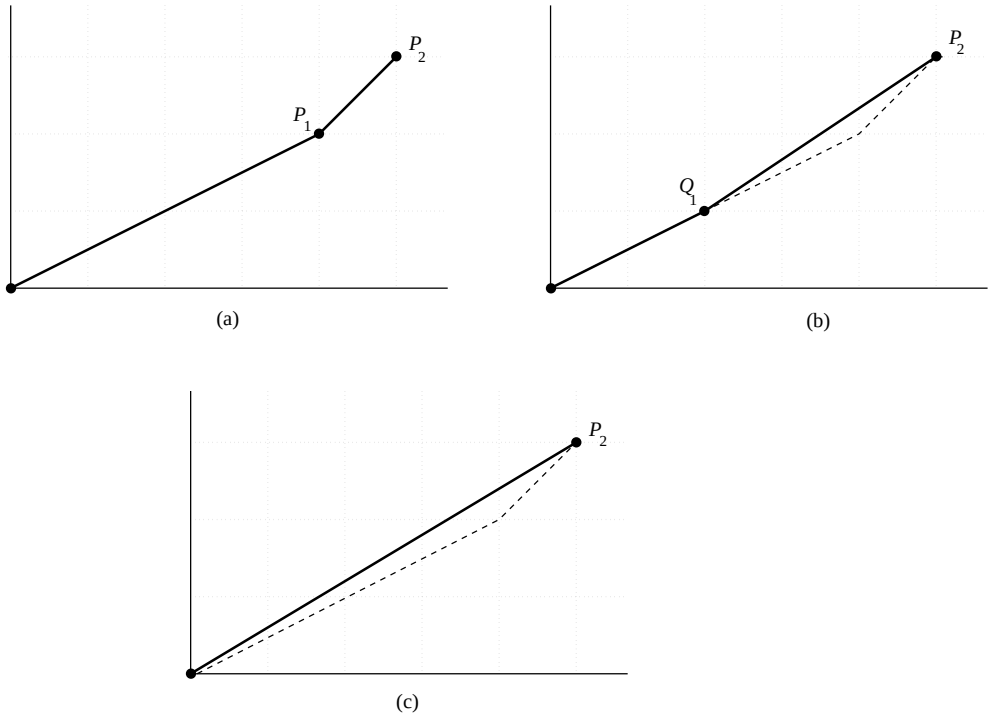


FIG. 3.4. *Newton diagrams corresponding to Example* 2. *In* (a) *the Newton diagram and envelope coincide. In* (b) *and* (c) *the diagram is shown as a solid line and the envelope as a dashed line.*

of ε in $\alpha_{k(1)}$ is different from zero, i.e., whether or not $\mathcal{Q}_1 = (2,1)$ is among the points plotted in the Newton diagram.

• If $\delta \neq 0$, then $a_{k(1)} = 1$ and there is a segment of slope $1/2$ in the Newton diagram connecting $(0,0)$ and $\mathcal{Q}_1$. If, additionally, $\det \Phi_2 \neq 0$, there is a second segment of slope $2/3$ between $\mathcal{Q}_1$ and $\mathcal{P}_2 = (5,3)$ (see Fig. 3.4 (b)). Hence, two of the eigenvalues are

$$\lambda_1^l(\varepsilon) = \delta^{1/2}\varepsilon^{1/2} + o(\varepsilon^{1/2})$$

since $\delta = \hat{\alpha}_{k(1)}$ is the unique nonzero eigenvalue of Φ_1. The other three eigenvalues of $A + \varepsilon B$ may be expanded as

$$\lambda_2^l(\varepsilon) = \eta^{1/3}\varepsilon^{2/3} + o(\varepsilon^{2/3}) \quad \text{for } \eta = \frac{\det \Phi_2}{\delta}$$

applying Corollary 3.3 (ii) (in this case $\alpha = n_2 = 1$ and $(\sigma n_1 - 1)\beta = 1/3 < 2/3 = \sigma$).

• If $\delta = 0$, then $a_{k(1)} > 1$ and $\mathcal{Q}_1$ does not appear in the diagram. In the case when $\det \Phi_2 \neq 0$, the Newton diagram consists of a single segment of slope $3/5$ connecting the origin with $\mathcal{P}_2$ (see Fig. 3.4(c)). The five eigenvalues are of the form

$$(\det \Phi_2)^{1/5}\varepsilon^{3/5} + o(\varepsilon^{3/5}).$$

Finally, if both δ and $\det \Phi_2$ are zero, the actual slopes of the Newton diagram depend on the vanishing of the four-dimensional minors of B.

4. Spectral condition numbers. The results of section 2 lead immediately to the proposition of a new notion of condition number for multiple eigenvalues.

DEFINITION 4.1. *Define the* Hölder condition number *of the eigenvalue* λ *by*

$$\mathrm{cond}(\lambda) = (n_1, \alpha),$$

where, as before, n_1 *is the dimension of the largest Jordan block associated with* λ *and*

$$\alpha = \max_{||B||\le 1} \mathrm{spr}(Y_1 B X_1),$$

where spr *denotes the spectral radius and the* r_1 *columns of* X_1 *(rows of* Y_1*) are independent right (left) eigenvectors of* λ*, each corresponding to a Jordan chain of greatest length* n_1 *as defined in section* 2.

The motivation for this definition is that $1/n_1$ is the smallest possible power of ε in the expansion of the eigenvalues of any perturbation $A + \varepsilon B$, while α^{1/n_1} is the largest possible magnitude of the coefficient of ε^{1/n_1} in such expansions. Clearly, it follows from Theorem 2.1 that for all $c > 1$, the eigenvalues λ' of $A + \varepsilon B$ converging to λ as $\varepsilon \downarrow 0$ satisfy

$$|\lambda' - \lambda| \le c\alpha^{1/n_1}\varepsilon^{1/n_1} \tag{4.1}$$

for all sufficiently small positive ε. In fact, this bound is sharp in the sense that given A, there exists a perturbation B such that for all $c < 1$, (4.1) holds with the inequality reversed for some perturbed eigenvalue λ' when ε is sufficiently small.

Note that the definition depends on the choice of matrix norm $||\cdot||$. We shall restrict our attention to unitarily invariant norms [6, p. 308].

THEOREM 4.2. *If the condition number* $\mathrm{cond}(\lambda) = (n_1, \alpha)$ *is defined by any unitarily invariant matrix norm* $||\cdot||$*, then*

$$\alpha = ||X_1 Y_1||_2.$$

Proof. One has

$$\begin{aligned}\max_{||B||\le 1} \mathrm{spr}(Y_1 B X_1) &= \max_{||B||\le 1} \mathrm{spr}(B X_1 Y_1) \\ &\le \max_{||B||\le 1} ||B X_1 Y_1||_2 \\ &\le ||X_1 Y_1||_2,\end{aligned}$$

where the final inequality follows because $||B||_2 \le ||B||$ for any unitarily invariant norm. To see that equality holds, note the following. Let the scalar σ_1, the row vector u_1, and the column vector v_1 be, respectively, the largest singular value and the corresponding left and right singular vectors of $X_1 Y_1$, so that $||u_1|| = ||v_1|| = 1$ and $u_1 X_1 Y_1 v_1 = \sigma_1 = ||X_1 Y_1||_2$. Setting $B = v_1 u_1$ gives $\mathrm{spr}(B X_1 Y_1) = \mathrm{spr}(u_1 X_1 Y_1 v_1) = \sigma_1$. □

Remarks.

1. In the case in which λ is simple, we have $n_1 = r_1 = 1$, so $\mathrm{cond}(\lambda) = (1, \alpha)$, where $\alpha = ||xy|| = ||x||\,||y||$, with the column vector x and the row vector y, respectively, right and left eigenvectors for λ, normalized so that $yx = 1$. Without loss of generality, one can take $||x|| = 1$, so $\alpha = ||y||$. Thus the condition number reduces to the standard definition [4, p. 152]; see also [5, p. 202], where the normalization used is $||x|| = ||y|| = 1$, giving the definition $1/(yx)$ for the condition number.

2. In the case in which λ is nonderogatory, we have $r_1 = 1$, so $\mathrm{cond}(\lambda) = (n_1, \alpha)$, where $\alpha = ||xy|| = ||x||\ ||y||$, with x and y, respectively, right and left eigenvectors. However, when $n_1 > 1$, we have $yx = 0$ (directly from the Jordan form). The normalization in this case is $Q_1^1 P_1^1 = I$, where the columns of P_1^1 (rows of Q_1^1) are right (left) Jordan chains for λ, x being the first column of P_1^1 and y the last row of Q_1^1. If A is in Jordan form, then $P_1^1 = Q_1^1 = I$, so $\mathrm{cond}(\lambda) = (n_1, 1)$.

For example, take

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

and consider the eigenvalue $\lambda = \delta$. For $\delta > 0$, one has $x = [1\ \ 0]^T$, $y = [1\ \ 1/(2\delta)]$, so $\mathrm{cond}(\delta) = (1, \alpha)$ with $\alpha = ||x||\ ||y|| = \sqrt{1 + 1/(4\delta^2)}$. Since the eigenvalue is simple, this condition number coincides with those given by [4] and [5]. As $\delta \downarrow 0$, the coefficient α in $\mathrm{cond}(\delta)$ diverges to $+\infty$. For $\delta = 0$, the eigenvalue $\lambda = \delta$ has multiplicity two, so the definitions given in [4, p. 152] and [5] do not apply. In this case, A is in Jordan form, so one has $x = [1\ \ 0]^T$, $y = [0\ \ 1]$, and $\mathrm{cond}(0) = (2, 1)$. Thus, although the condition number is not a continuous function of δ in the conventional sense, the divergence of α as $\delta \downarrow 0$ is reflected by the drop in the power $1/n_1$ at the limit point.

Chatelin [4, p. 156] also introduced a closely related Hölder condition number in the more general context of clusters of eigenvalues. Let us restrict our attention to the case in which the cluster consists of one multiple eigenvalue λ of multiplicity n. Chatelin defines a Hölder condition number $\mathrm{csp}(\lambda) = (n, \beta)$, with a coefficient β which depends on the conditioning of the transformation reducing the matrix to Jordan form. Specifically, consider the matrices P and Q in (2.1), let P have a "QR" factorization

$$P = UR, \qquad U^*U = I,$$

and define $V = RQ$. Thus, the columns of U form a unitary basis for the right invariant subspace for λ, while the rows of V form a (nonunitary) basis for the left invariant subspace, satisfying the normalization condition $VU = RQPR^{-1} = I$, since $QP = I$ from (2.2). Then the Chatelin condition number $\mathrm{csp}(\lambda) = (n_1, \beta)$ has β defined by

$$\beta = \mathrm{cond}_2(R)\, \|V\|_2,$$

where $\mathrm{cond}_2(R)$ is the ordinary condition number $\|R\|_2 \|R^{-1}\|_2$. (To see the equivalence with Chatelin's definition, note that $U^*AU = RJR^{-1}$, using $\widehat{Q}P = 0$, again from (2.2).)

An important advantage of $\mathrm{cond}(\lambda) = (n_1, \alpha)$ over $\mathrm{csp}(\lambda) = (n_1, \beta)$ is that the coefficient α depends only on the left and right *eigenvectors, not* on the Jordan vectors. As a consequence, we obtain the following relation between both condition numbers.

LEMMA 4.3. *If the condition number* $\mathrm{cond}(\lambda) = (n_1, \alpha)$ *is defined by any unitarily invariant matrix norm, then it is related to* $\mathrm{csp}(\lambda) = (n_1, \beta)$ *by*

$$\alpha \leq \beta.$$

Proof. First, note that

$$\alpha = \|X_1 Y_1\|_2 \leq \|X_1\|_2 \|Y_1\|_2 \leq \|P\|_2 \|Q\|_2$$

since the columns of X_1 (resp., rows of Y_1) are a subset of those of P (resp., Q). Now, $P = UR$ with $U^*U = I$ and $Q = R^{-1}V$, so

$$\alpha \leq \|P\|_2 \|Q\|_2 = \|R\|_2 \|R^{-1}V\|_2 \leq \mathrm{cond}_2(R)\, \|V\|_2 = \beta. \qquad \square$$

Consider, for example,

$$A = \begin{bmatrix} 0 & \delta \\ 0 & 0 \end{bmatrix}$$

with a double eigenvalue $\lambda = 0$. When $\delta > 0$, the Jordan form of A is given by

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = P^{-1} A P, \qquad P = \begin{bmatrix} 1 & 0 \\ 0 & 1/\delta \end{bmatrix}$$

with right eigenvector equal to the first column of P, i.e., $x = [1 \quad 0]^T$, and left eigenvector equal to the second row of P^{-1}, $y = [0 \quad \delta]$. Thus, for $\delta > 0$, $\mathrm{cond}(0) = (2, \alpha)$ with $\alpha = ||x||\ ||y|| = \delta$. The Chatelin condition number is $\mathrm{csp}(0) = (2, \beta)$ with $\beta = \max\{\delta, 1/\delta\}$, the condition number of P in the 2-norm. As $\delta \downarrow 0$, the coefficient α in $\mathrm{cond}(\lambda) = (2, \alpha)$ converges to zero, as it should since at the limit point the power $1/n_1$ *increases* to 1, giving the perfect condition number $\mathrm{cond}(0) = (1, 1)$. However, the coefficient β in $\mathrm{csp}(0) = (2, \beta)$ diverges to $+\infty$, even though the condition number in the limiting case $\delta = 0$ is also $\mathrm{csp}(0) = (1, 1)$.

The condition number $\mathrm{cond}(\lambda)$ is trivially extended to clusters of eigenvalues by defining it to be the lexicographic maximum of the ordered pairs defining the condition number for each element of the cluster.

Acknowledgments. The third author thanks Prof. A. Seyranian for bringing [13] to his attention several years ago and for suggesting the dedication of this paper. The authors also thank an anonymous referee for a number of helpful suggestions.

REFERENCES

[1] H. Baumgärtel, *Analytic Perturbation Theory for Matrices and Operators*, Birkhäuser-Verlag, Basel, Switzerland, 1985.
[2] J. V. Burke and M. L. Overton, *Stable perturbations of nonsymmetric matrices*, Linear Algebra Appl., 171 (1992), pp. 249–273.
[3] J. V. Burke and M. L. Overton, *Differential properties of the spectral abscissa and the spectral radius for analytic matrix-valued mappings*, Nonlinear Anal., 23 (1994), pp. 467–488.
[4] F. Chatelin, *Eigenvalues of Matrices*, John Wiley, New York, 1993.
[5] G. Golub and C. F. Van Loan, *Matrix Computations*, 1st ed., The Johns Hopkins University Press, Baltimore, MD, 1983.
[6] R. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1990.
[7] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1980.
[8] H. Langer and B. Najman, *Remarks on the perturbation of analytic matrix functions* III, Integral Equations Operator Theory, 15 (1992), pp. 796–806.

[9] H. Langer and B. Najman, *Leading coefficients of the eigenvalues of perturbed analytic matrix functions*, Integral Equations Operator Theory, 16 (1993), pp. 600–604.
[10] V. B. Lidskii, *Perturbation theory of non-conjugate operators*, U.S.S.R. Comput. Math. and Math. Phys., 1 (1965), pp. 73–85 (Zh. vychisl. Mat. mat. Fiz., 6 (1965), pp. 52–60).
[11] J. H. Maddocks and M. L. Overton, *Stability theory for dissipatively perturbed Hamiltonian systems*, Comm. Pure Appl. Math., 48 (1995), pp. 583–610.
[12] M. M. Vainberg and V. A. Trenogin, *Theory of Branching of Solutions of Non-linear Equations*, P. Noordhoff, Leyden, 1974.
[13] M. I. Vishik and L. A. Lyusternik, *The solution of some perturbation problems for matrices and selfadjoint or non-selfadjoint differential equations* I, Russian Math. Surveys, 15 (1960), pp. 1–74 (Uspekhi Mat. Nauk, 15 (1960), pp. 3–80).
[14] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon, Oxford, UK, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 4, pp. 818–826, October 1997

002

THE MATRIX DYNAMIC PROGRAMMING PROPERTY AND ITS IMPLICATIONS*

J. P. LE CADRE† AND O. TRÉMOIS‡

Abstract. The dynamic programming (DP) technique rests on a very simple idea, the principle of optimality due to Bellman. This principle is instrumental in solving numerous problems of optimal control. The control law minimizes a cost functional and is determined by using the optimality principle. However, applicability of the optimality principle requires that the cost functional satisfies the property called "matrix dynamic programming (MDP) property." A simple definition of this property will be provided and functionals having it will be considered.

Key words. dynamic programming, determinants, monotonicity, target tracking

AMS subject classifications. 49L20, 15A15, 15A45, 15A69

PII. S0895479895288334

1. Introduction.

The DP technique rests on a very simple idea, the principle of optimality due to Bellman [1]. This principle simply asserts that if $\pi^* = \{\mu_0^*, \mu_1^*, \ldots, \mu_n^*\}$ is an optimal control law then [1] the truncated control law $\{\mu_i^*, \mu_{i+1}^*, \ldots, \mu_n^*\}$ is optimal for the ith truncated control problem.

This principle is instrumental in solving numerous problems of optimal control of a dynamic system over a finite number of stages (finite horizon). We refer to [1, 2] for a thorough and motivated presentation of the DP technique.

The control law (or the strategy [2]) must minimize a cost functional. However, to our knowledge, the authors always assume that the cost functional is additive over time.

The problem we deal with consists of optimizing the trajectory of a passive receiver. For practical purposes, we must minimize a functional depending on the state values and the control law. The functional is a functional of the Fisher information matrix (FIM) since, roughly speaking, the FIM is a general measure of the estimability problem [3, 4, 5]. A general presentation of our problem is given in [6, 7, 8].

Various choices of the FIM functional have been considered in the literature [9], even if both theoretical and practical considerations advocate for the use of the determinant [8, 10, 11]. At this point, it is necessary to stress that the determinant is not linear so the cost functional additivity no longer holds. Actually, applicability of the principle of optimality to the matrix case requires that the cost functional satisfies to the matrix dynamic programming (MDP) property. A simple definition of the MDP property will be provided, and we shall examine the functionals having it.

Then, it is shown that they are reduced to the functionals of the form $f(A) = g(\mathrm{tr}(AM))$ (cf. Proposition 2.2). Consequently, these functionals are "almost" linear (obviously a linear functional yields an additive cost), which may be rather restrictive.

At this point, it is worth recalling the special structure of the FIMs. Actually, if we restrict our attention to a very specific estimation problem (namely, target motion

* Received by the editors June 26, 1995; accepted for publication (in revised form) by G. Cybenko September 6, 1996. This work has been supported by DCN /Ing/Sud (Direction des Constructions Navales), France.

http://www.siam.org/journals/simax/18-4/28833

† IRISA/CNRS, Campus de Beaulieu, 35042 Rennes cedex, France (lecadre@irisa.fr).

‡ Thomson-Marconi Sonar, Route de Sainte-Anne du Portzic, Brest, France (olivier.o.t.tremois@tms.thomson.fr).

analysis (TMA)), which deals with the estimation of the kinematic parameters defining a source trajectory, then the FIM matrices exhibit a very special structure [12] which is very succinctly presented in the appendix. We shall then consider the applicability of the optimality principle to this class of matrices for the det functional. Some results are thus obtained, but they cannot be extended to the generic case.

Throughout the text, the following notations will be used:

- a capital letter denotes a matrix,
- a capital calligraphic letter denotes a subspace,
- the symbol (*) means transposition conjugation,
- the symbols det and tr stand, resp., for the determinant and the trace,
- $\mathcal{H}_n$ is the space of n-dimensional Hermitian matrices,
- $\mathcal{P}_n$ (resp., $\mathcal{P}_n^+$) is the subset of positive semidefinite (resp., positive definite) matrices,
- I is the identity matrix,
- $A \succeq B$ means that the matrix $A - B$ is semidefinite positive.

The paper is organized as follows. The MDP property is introduced in section 2. General results are then obtained. The validity of the optimality principle for the determinant of structured matrices is considered in section 3.

2. The MDP property and its implications. We shall say that the functional f defined from $\mathcal{H}_n$ (the vector space of n-dimensional Hermitian matrix) to $\mathbb{R}$ satisfies the MDP property if the following conditions are fulfilled.

DEFINITION 2.1.

- *f is smooth* $\left(\mathcal{C}^2\right)$,
- *let A and B in $\mathcal{H}_n$ be two matrices and assume that $f(B) > f(A)$; then whatever the matrix C in $\mathcal{H}_n$, we have $f(B + C) > f(A + C)$.*

An interpretation of this definition in terms of dynamic programming is the following type of inequality [6, 7]:[1]

$$\min f\left\{\sum_{j\in S}[C_{i,j}(d) + F_{\pi_1^*}(k+1,j)]p_{i,j}(d)\right\} \leq f\left\{\sum_{j\in S}[C_{i,j}(d) + F_{\pi_1}(k+1,j)]p_{i,j}(d)\right\},$$

which must be valid for the strategy π_1^*, optimal up to $k+1$.

The question we deal with consists of determining the functionals f having the MDP property. An answer to this question is provided with the following result.

PROPOSITION 2.2. *Let f satisfy the MDP property; then*

$$f(A) = g\left(\operatorname{tr}(AR)\right),$$

where g is any monotone-increasing function and R is a fixed matrix.

Proof. Since it has been assumed that f is smooth, we can consider the first-order expansion[2] of f around A

$$f\left(A + \rho\, C\right) \stackrel{1}{=} f(A) + \rho \operatorname{tr}[\nabla^* f(A)\, C] + 0(\rho) \tag{1}$$

(ρ scalar).

[1] F denotes an FIM matrix.

[2] The symbol $\stackrel{1}{=}$ denotes a first-order expansion.

In the above formula, $\nabla f(A)$ denotes the gradient vector of f in A. Actually, the notation $\mathrm{tr}[\nabla^* f(A)\ C]$ replaces the true expression [13, 16] of the differential of $f, Df_A(C)$ and corresponds to (see comments)

$$Df_A(C) = \mathrm{tr}[\nabla^* f(A)\ C]. \tag{2}$$

Assume now that the gradient vectors $\nabla f(A)$ are not colinear altogether. Then there exist two matrices A and B s.t. $\nabla f(A) \neq \nabla f(B)$. Denoting $F^{\perp}$ as the subspace orthogonal to F (for the classical scalar matrix product [14]), we thus have

$$(P_1) \quad (\nabla f(A))^{\perp} \neq (\nabla f(B))^{\perp}. \tag{3}$$

At this point, stress that the matrices A and B satisfying (P_1) may be chosen as close (for the Frobenius norm [15]) as we want.

As a consequence of (P_1) there exists a matrix C such that

$$\mathrm{tr}[\nabla^* f(A)\ C] \neq 0 \ \text{ and } \ \mathrm{tr}[\nabla^* f(B)\ C] = 0.$$

If $\mathrm{tr}[\nabla^* f(A)\ C] < 0$, then $\mathrm{tr}[\nabla^* f(A)\ (-C)] \ > 0$, so we can assume that

$$\mathrm{tr}[\nabla^* f(A)\ C] \ > 0 \text{ and } \mathrm{tr}[\nabla^* f(B)\ C] = 0. \tag{4}$$

Now consider the function $g(\rho)$

$$g(\rho) \stackrel{\Delta}{=} f\,(B + \rho\ C) - f\,(A + \rho\ C)$$

and its first-order expansion around 0, i.e.,

$$g(\rho) = f(B) - f(A) - \rho\ \mathrm{tr}[\nabla^* f(A)\ C] + 0(\rho). \tag{5}$$

Since the functional f is continuous on $\mathcal{H}_n$, we may choose (A, B) such that

$$f(B) - f(A) = \frac{\rho}{2}\mathrm{tr}[\nabla^* f(A)\ C],$$

and, consequently,

$$\begin{aligned} &f(B + \rho\ C) - f(A + \rho\ C) = -\frac{\rho}{2}\ \mathrm{tr}[\nabla^* f(A)\ C] + 0(\rho) \\ &(\mathrm{tr}[\nabla^* f(A)\ C] \ > 0)\,. \end{aligned} \tag{6}$$

The above equality implies that f does not satisfy the MDP property.

Therefore, if f has the MDP property then all its gradients are colinear and proportional to a unique vector. Denote this vector by $\mathbf{G}$; we thus have

$$\begin{aligned} &\nabla f(A) = \lambda(A)\mathbf{G} \quad \forall A \in \mathcal{H}_n \\ &(\lambda(A) \text{ scalar })\,. \end{aligned} \tag{7}$$

Now if we recall the intermediate value theorem and the differentiation chain rule [16]

$$\nabla g\,[h(A)] = g'\,(h(A))\,\nabla h(A)$$

$$(g : \mathbb{R} \longrightarrow \mathbb{R}, h : \mathcal{H}_n \longrightarrow \mathbb{R})\,,$$

we conclude that f is the composition of a scalar function g and a linear form h. Such a linear form h may always be written $h(A) = \mathrm{tr}(AR)$, where R is a fixed matrix.

Reciprocally, it is a trivial matter to show that $f(A) = g\,(\mathrm{tr}(AR))$ with a g monotone increasing function that satisfies the MDP property. The proof is thus complete. □

Comments.

1. Consider for instance $f(A) = \log \det A$; then [14]

$$Df_A(C) = \operatorname{tr}\left(A^{-1}C\right) = \operatorname{tr}[\nabla^* f(A)\, C]$$

(A invertible),
and we see immediately that f does not have the MDP property.

2. The same remark is valid for functionals as simple as $f(A) = \operatorname{tr}\left(A^{-1}\right)$.

But actually we only need the following condition on f.

DEFINITION 2.3. *The functional f has the MDP*1 *property if the following conditions are satisfied.*

For all positive definite matrices A and B and positive semidefinite matrix C, the following property holds:

$$f(B) > f(A) \Rightarrow f(B + C) > f(A + C) \tag{8}$$

$\left(f : \mathcal{C}^2\right)$.

Note that MDP1 is a refinement of the MDP property and may, possibly, be less demanding than MDP. At this point, it is worth mentioning the following lower bound of $\det(A + B)$. We refer, for instance, to [17, pp. 229–230, 18] for a proof.

LEMMA 2.4. *Let A, B be Hermitian $n \times n$ matrices, and suppose that A is positive definite and that B is nonnegative definite. Then*

$$\det(A + B) \geq \det(A) + \frac{\det(A)}{n\lambda_{\max}} \operatorname{tr}(B). \tag{9}$$

Here $\lambda_{\max}$ denotes the maximum eigenvalue of A.

However, the following counterexample shows that MDP1 is not satisfied by the "det" functional.

Counterexample.

$$A_\varepsilon = \begin{pmatrix} 1+\varepsilon & 0 \\ 0 & 5 \end{pmatrix}, \quad B = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix},$$

$$\det(B) = 5, \det(A_\varepsilon) = 5(1+\varepsilon), A_\varepsilon \text{ and } B \succ 0 \text{ (} \varepsilon \text{ suff. small)},$$

$$C = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

and thus

$$\det(A_\varepsilon + C) = 11 + 6\varepsilon \text{ and } \det(B + C) = 7.$$

It is quite obvious that this (elementary) counterexample is not restricted to a rank deficient C matrix since C may be slightly perturbed without changing the sign of $\det(B + C) - \det(A + C)$. Therefore for ε sufficiently small (and negative) we have

$$\det(B) > \det(A_\varepsilon) \text{ and } \det(A_\varepsilon + C) > \det(B + C).$$

Similarly, we can show that MDP1 is not satisfied by the functional $f(A) = \operatorname{tr}\left(A^{-1}\right)$ but is trivially satisfied by any functional $f(A) = g(\operatorname{tr}(AR))$ (g: monotone increasing). We thus consider the following problem.

What are the functionals satisfying the MDP1? We may reasonably suspect that they are reduced to those satisfying the MDP. However, the proof of Proposition 2.2 cannot be trivially extended since the subset of semidefinite matrices is not a subspace. Actually, if C is in $\mathcal{P}_n^+$ then $-C$ is not in $\mathcal{P}_n^+$ and the reasoning leading to (4) is not valid. The difficulty comes from the fact that $\mathcal{P}_n^+$ is a convex subset of $\mathcal{H}_n$ and not a subspace.

3. Structured determinants and the MDP property. Our attention will now be restricted to structured matrices. Various structures will be considered corresponding to various scenarios (see [6]) of target motion analysis. The statistical motivations of such special matrix structures are beyond the scope of this paper, but the true problems are thus conveniently described [12, 7].

Since the general MDP property is not satisfied by the determinant, we shall consider special cases associated with particular matrix structures and specific definitions of the "addition." It will then be shown that the DP property may be extended, but the validity of these extensions is limited.

3.0.1. One-leg scenario. In this case, the elementary FIM $F(A,C)$ takes the following form:

$$F(A,C) = \begin{pmatrix} A & iA \\ iA & i^2A \end{pmatrix} + \begin{pmatrix} C & jC \\ jC & j^2C \end{pmatrix} = \begin{pmatrix} A+C & iA+jC \\ iA+jC & i^2A+j^2C \end{pmatrix}$$

with

$$A, C \in \mathcal{P}_n^+,\ i, j \in \mathbb{N}.$$

Then, the following result holds.

PROPOSITION 3.1.

$$\det\left(F(A,C)\right) = (j-i)^{2n} \det A\ \det C.$$

Proof. An elementary proof relies on the following factorization:

$$F = \begin{pmatrix} A & I \\ iA & jI \end{pmatrix} \begin{pmatrix} I & iI \\ C & jC \end{pmatrix},$$

hence

$$\det F = \det \begin{pmatrix} A & I \\ iA & jI \end{pmatrix} \det \begin{pmatrix} I & iI \\ C & jC \end{pmatrix}. \tag{10}$$

Now using the classical lemma about the determinant of a partitioned matrix [15, 19] we obtain directly

$$\begin{aligned} \det \begin{pmatrix} A & I \\ iA & jI \end{pmatrix} &= \det A \det\left(jI - iAA^{-1}\right) \\ &= \det A \det\left((j-i)I\right) \end{aligned}$$

and similarly

$$\det \begin{pmatrix} I & iI \\ C & jC \end{pmatrix} = \det C\ \det\left((j-i)I\right). \tag{11}$$

□

Using the proof of Proposition 3.2 (see below), Proposition 3.1 may be extended to the case $A \in \mathcal{P}_n^+$, $C \in \mathcal{P}_n$.

A direct consequence of Proposition 3.1 is

$$\det B > \det A > 0 \Rightarrow \det F(B,C) > \det F(A,C) \quad (C \succeq 0). \tag{12}$$

Actually, as we shall see later, the simplicity of $\mathbf{Res}_1$ is a consequence of the rank deficiency of the following matrix:

$$\begin{pmatrix} A & iA \\ iA & i^2A \end{pmatrix} = \begin{pmatrix} 1 & i \\ i & i^2 \end{pmatrix} \otimes A$$

($\otimes$: Kronecker product [15]).

Thus, the MDP property is verified for this particular matrix structure. However, for practical applications, Proposition 3.1 should be extended to the following two problems.

3.0.2. Problem 1. In fact, statistical considerations may lead us to consider a slightly more general structure

$$\begin{pmatrix} A & \alpha A \\ \alpha A & \beta A \end{pmatrix}.$$

This matrix is no longer rank deficient (in general) so that previous calculations are not valid. However, the following result holds.

PROPOSITION 3.2.

$$\det\left[\begin{pmatrix} A & \alpha A \\ \alpha A & \beta A \end{pmatrix} + \begin{pmatrix} C & jC \\ jC & j^2C \end{pmatrix}\right] = \det A \det\left[\left(\beta - \alpha^2\right) A + \left(\beta - 2j\alpha + j^2\right) C\right].$$

Proof. That A is positive definite admits a decomposition in triangular factors ($A = TT^*$) so that

$$\begin{pmatrix} A+C & \alpha A + jC \\ \alpha A + jC & \beta A + j^2 C \end{pmatrix} = \begin{pmatrix} T & 0 \\ 0 & T \end{pmatrix} \begin{pmatrix} I+S & \alpha I + jS \\ \alpha I + jS & \beta I + j^2 S \end{pmatrix} \begin{pmatrix} T^* & 0 \\ 0 & T^* \end{pmatrix}$$

with

$$S \triangleq T^{-1} C T^{-1*}. \tag{13}$$

Now

$$\det \begin{pmatrix} I+S & \alpha I + jS \\ \alpha I + jS & \beta I + j^2 S \end{pmatrix} = \det\left[(I+S)\left(\beta I + j^2 S\right) - (\alpha I + jS)^2\right]$$

(where we have used the fact that $(I+S)$ and $(\alpha I + jS)$ commute)

$$= \det\left(\left(\beta - \alpha^2\right) I + \left(\beta - 2j\alpha + j^2\right) S\right), \tag{14}$$

which ends the proof. □

If $\beta = \alpha^2$, then we have

$$\det F(A,C) = \det A \det\left[(\alpha - j)^2 C\right]$$

and therefore

$$\det B > \det A \Rightarrow \det F(B,C) > \det F(A,C) \quad (C \succeq 0). \tag{15}$$

If $\beta \neq \alpha^2$, then the previous implication does not hold and the MDP property is not valid.

3.0.3. Problem 2. Another important problem arises when we try to extend the previous calculations to more than two matrices, i.e., to calculate the following structured determinant:

$$\det \begin{pmatrix} A + C_1 + C_2 & A + jC_1 + kC_2 \\ A + jC_1 + kC_2 & A + j^2C_1 + k^2C_2 \end{pmatrix}. \tag{16}$$

The previous simple results (Proposition 3.1 or 3.2) cannot be extended to this structure. It seems impossible to obtain a simple expression of this determinant even when using more sophisticated algebra [20]. This constitutes a major problem. A last example considers a very special addition law where the dimension of the matrices is increasing. As previously noted, this special structure may be justified by statistical considerations [6, 7, 8] associated with multileg scenarios.

3.1. Multileg scenarios. Considering this type of scenario leads to increasing the dimensionality of the state vector and thus to considering the following elementary structure of the matrix F:

$$\mathcal{F}(A, C) = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} C & iC \\ iC & i^2C \end{pmatrix},$$

$$A \in \mathcal{P}_n^+,\ C \in \mathcal{P}_n^+,\ i \in \mathbb{N}.$$

We then obtain the following result.

PROPOSITION 3.3.

$$\det \mathcal{F}(A, C) = (i)^{2n} \det A \det C.$$

Proof.

$$\begin{aligned} \det\left(\mathcal{F}(A, C)\right) &= \det\left[i^2C - i^2C(A + C)^{-1}C\right] \det(A + C) \\ &= i^{2n}(\det C)^2 \det\left[C^{-1}(A + C) - I\right] \\ &= i^{2n} \det C \det A. \qquad \square \end{aligned} \tag{17}$$

In view of Proposition 3.3, the following property holds:

$$\det B > \det A \Rightarrow \det \mathcal{F}(B, C) > \det \mathcal{F}(A, C). \tag{18}$$

4. Conclusion. Applicability of the principle of optimality to matrix cost functionals requires that the MDP property be satisfied. A simple definition of this property has been given, and we have determined the functionals that have it. Since the det functional does not satisfy the MDP property, various special structures have been considered.

5. Appendix. The aim of this appendix is to provide a very succinct presentation of the calculation of the FIM matrices in the TMA context. For more details, we refer to [8, 12]. First consider a rectilinear and uniform motion of the source. The source, located at the coordinates (r_{xs}, r_{ys}), moves with a constant velocity vector $\mathbf{v}\,(v_{xs}, v_{ys})$ and is thus defined to have the state vector

$$\mathbf{X}_s \triangleq [r_{xs}, r_{ys}, v_{xs}, v_{ys}]^* \,. \tag{19}$$

The receiver state is similarly defined as

$$\mathbf{X}_{rec} \triangleq [r_{x\ rec}, r_{y\ rec}, v_{x\ rec}, v_{y\ rec}]^*$$

so that, in terms of the relative state vector $\mathbf{X}$ defined by

$$\mathbf{X} = \mathbf{X}_s - \mathbf{X}_{rec} \triangleq [r_x, r_y, v_x, v_y]^*,$$

the discrete time equation of the system (i.e., the equation of the relative motion) takes the following form:

$$\mathbf{X}_{k+1} = F\mathbf{X}_k + \mathbf{U}_k,$$

where

$$F = \Phi\,(k, k+1) = \begin{pmatrix} Id & \alpha Id \\ 0 & Id \end{pmatrix},\ Id \triangleq \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and

$$\alpha \triangleq t_{k+1} - t_k = cst\ . \tag{20}$$

The measurement noise w_k is usually modelled by an independently and identically distributed (i.i.d.) zero-mean, Gaussian process.

The partial derivative matrix of the bearing vector $\mathbf{\Theta}(\mathbf{X})$ with respect to the state components is easily calculated [12] yielding

$$\frac{\partial \mathbf{\Theta}(\mathbf{X})}{\partial \mathbf{X}} = \begin{pmatrix} \frac{\cos\theta_1}{r_1} & -\frac{\sin\theta_1}{r_1} & \frac{\cos\theta_1}{r_1} & -\frac{\sin\theta_1}{r_1} \\ \vdots & & & \\ \frac{\cos\theta_n}{r_n} & -\frac{\sin\theta_n}{r_n} & \frac{n\cos\theta_1}{r_n} & -\frac{n\sin\theta_n}{r_n} \end{pmatrix},$$

where $\{\theta_i\}$ represents the source bearing (angle) at the instant i, and $\{r_i\}$ is the source-receiver distance.

Consider the case of a nonmaneuvering source (constant-velocity vector); then the calculation of the FIM is a routine exercise yielding [12]

$$\text{FIM} = \left(\frac{\partial \mathbf{\Theta}(\mathbf{X})}{\partial \mathbf{X}}\right)^* \Sigma^{-1} \left(\frac{\partial \mathbf{\Theta}(\mathbf{X})}{\partial \mathbf{X}}\right) \tag{21}$$

$$\triangleq \begin{pmatrix} \sum_{i=1}^n \Omega_i & \sum_{i=1}^n i\Omega_i \\ \sum_{i=1}^n i\Omega_i & \sum_{i=1}^n i^2\Omega_i \end{pmatrix}.$$

A realistic assumption consists of modelling the source trajectory by a sequence of elementary rectilinear uniform motions (named "legs"). The previous calculation of the FIM may be extended to this modelling, and the FIM then takes the following form [7, 6] (l legs):

$$\text{FIM} = \sum_{m=1}^{l} \sum_{k=(m-1)+j}^{mj} [\mathbf{d}_{m-1,l+1}(k)\mathbf{d}_{m-1,l+1}(k)^*] \otimes \Omega_k,$$

where $\mathbf{d}$ is a vector describing the index leg, consisting of 0 and 1, and Ω_k is a 2×2 elementary FIM.

REFERENCES

[1] D. P. Bertzekas, *Dynamic Programming, Deterministic and Stochastic Models*, Prentice–Hall, Englewood Cliffs, NJ, 1987.

[2] S. M. Ross, *Introduction to Stochastic Dynamic Programming*, Academic Press, New York, 1983.

[3] H. V. Poor, *An Introduction to Signal Detection and Estimation*, Springer-Verlag, New York, 1988.

[4] Y. Baram and T. Kailath, *Estimability and regulability of linear systems*, IEEE Trans. Automat. Control, 33 (1988), pp. 1116–1121.

[5] M. Segal and E. Weinstein, *A new method for evaluating the log-likelihood gradient, the Hessian and the Fisher information matrix for linear dynamic systems*, IEEE Trans. Inform. Theory, 35 (1989), pp. 682–687.

[6] J. P. Le Cadre and O. Trémois, *Optimization of the observer motion using dynamic programming*, in Proc. IEEE 1995, Internat. Conf. Acoustics, Speech Signal Proc., Vol. 5, Detroit, MI, 1995, pp. 3567–3570.

[7] O. Trémois, *Etude de méthodes de trajectographie pour des sources manoeuvrantes*, Ph.D. thesis, université de Rennes I, Rennes, France, 1995.

[8] J. P. Le Cadre and C. Jauffret, *Discrete-time observability and estimability analysis for bearings-only target motion analysis*, IEEE Trans. Aerospace Electron. Systems, 33 (1997), pp. 178–201.

[9] D. Ucinski, J. Korbicz, and M. Zaremba, *On optimization of sensor motions in parameter identification of two-dimensional distributed systems*, in Proc. European Control Conference 93, Grenoble, France, 1993, pp. 1359–1364.

[10] P. T. Liu, *An optimum approach in target tracking with bearing measurements*, J. Optim. Theory Appl., 56 (1988), pp. 205–214.

[11] S. E. Hammel and V. J. Aidala, *Observability requirements for three-dimensional tracking via angle measurements*, IEEE Trans. Aerospace Electron. Systems, 21 (1985), pp. 200–207.

[12] S. C. Nardone, A. G. Lindgren, and K. F. Gong, *Fundamental properties and performance of conventional bearings-only target motion analysis*, IEEE Trans. Automat. Control, 29 (1984), pp. 775–787.

[13] D. S. G. Pollock, *Tensor products and matrix differential calculus*, Linear Algebra Appl., 67 (1985), pp. 169–193.

[14] W. H. Greub, *Linear Algebra*, 4th ed., Springer-Verlag, New York, 1976.

[15] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, London, 1985.

[16] H. Cartan, *Calcul Différentiel*, Hermann, Paris, 1967.

[17] M. H. A. Davis and R. B. Vinter, *Stochastic Modelling and Control*, Monographs Statist. Appl. Probab., Chapman and Hall, London, 1985.

[18] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*, Mathematics in Science and Engineering 143, Academic Press, New York, 1979.

[19] R. A. Brualdi and H. Schneider, *Determinental identities: Gauss, Schur, Sylvester, etc.*, Linear Algebra Appl., 52–53 (1983), pp. 769–791.

[20] M. Marcus, *Finite Dimensional Multilinear Algebra*, Parts I and II, Marcel Dekker, New York, 1973 and 1975.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 4, pp. 827–841, October 1997

003

DISTANCES IN WEIGHTED TREES AND GROUP INVERSE OF LAPLACIAN MATRICES*

STEPHEN J. KIRKLAND†, MICHAEL NEUMANN‡, AND BRYAN L. SHADER§

Abstract. In this paper we find formulas for group inverses of Laplacians of weighted trees. We then develop a relationship between entries of the group inverse and various distance functions on trees. In particular, we show that the maximal and minimal entries on the diagonal of the group inverse correspond to certain pendant vertices of the tree and to a centroid of the tree, respectively. We also give a characterization for the group inverses of the Laplacian of an unweighted tree to be an M-matrix.

Key words. Laplacian matrix, generalized inverse, weighted tree

AMS subject classifications. Primary, 15A09; Secondary, 05C50

PII. S0895479896298713

1. Introduction. In this paper, for a weighted tree on n-vertices, we bring into focus the relationship between the inverse weighted generalized distance function which is defined on its vertices and the entries of the group inverse of the Laplacian matrix associated with the tree. In so doing we generalize and extend results which we have obtained in a previous paper for unweighted trees. We also shift the frame of reference for what we termed in that paper as *bottleneck numbers for the tree* to the inverse weighted distance function. Other results follow.

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 to vertex i, and whose (i, j)th off-diagonal entry equals 0 if there is no edge joining vertices i and j and equals the negative of the weight of the edge joining vertices i and j otherwise.

Suppose now that $\mathcal{G}$ 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

$$\tilde{d}(i,j) = \sum_{e \in \mathcal{P}_{i,j}} \frac{1}{w(e)}; \tag{1.1}$$

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*

*Received by the editors February 12, 1996; accepted for publication (in revised form) by R. Brualdi September 13, 1996.

http://www.siam.org/journals/simax/18-4/29871.html

†Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan, Canada S4S 0A2 (kirkland@max.cc.uregina.ca). This author's research was supported by NSERC grant OGP0138251.

‡Department of Mathematics, University of Connecticut, Storrs, CT 06269-3009 (neumann@math.uconn.edu). This author's research was supported by NSF grant DMS-9306357.

§Department of Mathematics, University of Wyoming, Laramie, WY 82071 (bshader@uwyo.edu). This author's research was partially supported by NSA grant MDA904-94-H-2051.

status of vertex i as the sum

$$\tilde{d}_i = \sum_{u \in T} \tilde{d}(u,i). \tag{1.2}$$

Our terminology follows that of Harary [6] for unweighted graphs.

Recall that for an $n \times n$ matrix A, the *group inverse of* A, when it exists, is the unique $n \times n$ matrix satisfying the matrix equations

$$\text{(i) } AXA = A, \text{ (ii) } XAX = X, \text{ and (iii) } AX = XA.$$

It is known that for real square matrix, A, the group inverse exists if and only if the Jordan blocks of A corresponding to the eigenvalue $\lambda = 0$ are all 1×1. It is customary to denote the group inverse of A by $A^{\#}$.

Now let $\mathcal{G}$ be a weighted tree on n vertices. It is readily seen from the definition of the Laplacian of $\mathcal{G}$ that L is a symmetric matrix with nonpositive off-diagonal entries and zero row sums. It follows now from the Perron–Frobenius theory, see, for example, Berman and Plemmons [1], that L is a positive semidefinite and irreducible M-matrix. In particular, the group inverse $L^{\#}$ of L exists.

In section 3 we shall show that for some constant c,

$$\left(L^{\#}_{1,1}, \ldots, L^{\#}_{n,n}\right) = \frac{1}{n}(\tilde{d}_1, \ldots, \tilde{d}_n) + c\mathbf{1},$$

where $\mathbf{1}$ is the n-vector of all 1's. Thus the maximal and minimal diagonal entries in $L^{\#}$ correspond to the vertices of maximal and minimal inverse status, respectively. We shall further show that of necessity such vertices must be pendant and centroid vertices of the tree, respectively. We also find a representation in terms of the inverse status values for the off-diagonal entries of $L^{\#}$ and show that its entries corresponding to the edges along the same path emanating at any vertex decrease as we move away from the vertex. This yields a characterization for $L^{\#}$ to be an M-matrix itself, and we analyze the unweighted trees whose Laplacian has a group inverse which is such a matrix.

The development of distance formulas for the entries of the group inverse of the Laplacian requires the preparation of some preliminary results for weighted trees of certain parameters associated with the graph which, for unweighted trees, we called *bottleneck numbers*. In section 2 we extend the notion and results on bottleneck numbers in [7] to weighted graphs.

2. Formulas for $L^{\#}$. Recall that the Laplacian matrix of a weighted connected graph is an irreducible singular M-matrix. In this section we establish general formulas for the group inverse of an irreducible singular M-matrix and then give combinatorial descriptions in the case of weighted graphs.

We begin with the following block matrix description. Let A be an $n \times n$ irreducible, singular M-matrix. Then there exists a positive vector x such that $Ax = 0$. The vector x is called a *right null vector* of A. Similarly, a *left null vector* of A is a positive vector y such that $y^T A = 0^T$. We also note, for $k = 1, 2, \ldots, n$, that the principle submatrix $A[\overline{\{k\}}, \overline{\{k\}}]$ obtained from A by deleting its kth row and column is a nonsingular M-matrix. The proof of the following is similar to that of Theorem 8.5.2 in [2].

PROPOSITION 2.1. *Let A be an irreducible singular M-matrix with right null*

vector $x = (x_1, x_2, \ldots, x_n)^T$ *and left null vector* $y = (y_1, y_2, \ldots, y_n)^T$. *Then*

$$A^{\#} = \frac{\hat{y}^T M \hat{x}}{(y^T x)^2} xy^T + \left[\begin{array}{c|c} M - \frac{1}{y^T x} M\hat{x}\hat{y}^T - \frac{1}{y^T x}\hat{x}\hat{y}^T M & \frac{-y_n}{y^T x} M\hat{x} \\ \hline \frac{-x_n}{y^T x}\hat{y}^T M & 0 \end{array}\right],$$

where $\hat{x} = (x_1, x_2, \ldots, x_{n-1})^T$, $\hat{y} = (y_1, y_2, \ldots, y_{n-1})^T$, *and* $M = A[\overline{\{n\}}, \overline{\{n\}}]^{-1}$.

If A is an $n \times n$ irreducible singular M-matrix, we call the nonnegative matrix $A[\overline{\{i\}}, \overline{\{i\}}]^{-1}$ the *bottleneck matrix* of A *based at* i. Proposition 2.1 describes $A^{\#}$ in terms of the bottleneck matrix of A based at n. If A is the Laplacian matrix of a weighted graph $\mathcal{G}$, then we also refer to $A[\overline{\{i\}}, \overline{\{i\}}]^{-1}$ as the *bottleneck matrix of* $\mathcal{G}$ *based at vertex* i. The following is an immediate consequence of Proposition 2.1 and the fact that $\mathbf{1}$ is a left and right null vector of the Laplacian matrix of a weighted graph.

PROPOSITION 2.2. *Let* L *be the Laplacian matrix of a connected weighted graph with* n *vertices. Then*

$$L^{\#} = \frac{\mathbf{1}^T M \mathbf{1}}{n^2} J + \left[\begin{array}{c|c} M - \frac{1}{n}MJ - \frac{1}{n}JM & -\frac{1}{n}M\mathbf{1} \\ \hline -\frac{1}{n}\mathbf{1}^T M & 0 \end{array}\right],$$

where $M = L[\overline{\{n\}}, \overline{\{n\}}]^{-1}$ *is the bottleneck matrix based at vertex* n.

If A is an $m \times n$ matrix, the submatrix of A whose rows have index in α and whose columns have index in β is denoted by $A[\alpha, \beta]$. We use $\overline{\alpha}$ and $\overline{\beta}$ to denote the complement of α in $\{1, 2, \ldots, m\}$ and of β in $\{1, 2, \ldots, n\}$, respectively.

We now develop a combinatorial description of the entries of the bottleneck matrix of a weighted graph. Let L be the Laplacian matrix of a connected weighted graph $\mathcal{G}$ with vertices 1, 2, ..., n. By the cofactor formula for the inverse, the (i, j)th entry of the bottleneck matrix of L based at n equals

$$\frac{(-1)^{i+j} \det L[\overline{\{j, n\}}, \overline{\{i, n\}}]}{\det L[\overline{\{n\}}, \overline{\{n\}}]}. \tag{2.1}$$

Our description of the bottleneck matrix follows from (2.1) and a generalization of the matrix-tree theorem obtained by Chaiken in [3]. The generalization is proven using the Cauchy–Binet formulas and gives a combinatorial description of the determinants of the square submatrices of a Laplacian matrix.

For the purpose of achieving the above, we need a few further graph theoretical notions. A *subgraph* of $\mathcal{G}$ is a graph $\mathcal{H}$ whose vertices are a subset of $1, 2, \ldots, n$, and whose edges are a subset of those of $\mathcal{G}$. If $\mathcal{G}$ is a weighted graph and E is a subset of edges of $\mathcal{G}$, then the *weight* of E is denoted by $w(e)$ and is the product of the weights of the edges in E. The *weight* of a subgraph $\mathcal{H}$ is the weight of its set of edges and the weight of a graph with no edges is defined to be 1. The set of all spanning trees of $\mathcal{G}$ is denoted by $\mathcal{S}$. We now define a special type of spanning forest. Let i, j, and k be (not necessarily distinct) vertices of $\mathcal{G}$. An $(\{i, j\}, k)$*-spanning forest* of $\mathcal{G}$ is a spanning forest of $\mathcal{G}$ which has exactly two connected components, one of which contains vertex k and the other of which contains the vertices i and j. The set of all $(\{i, j\}, k)$-spanning forests of $\mathcal{G}$ is denoted by $\mathcal{S}_k^{\{i,j\}}$.

Chaiken's all minors matrix tree theorem in [3] implies that

$$\det L[\overline{\{n\}}, \overline{\{n\}}] = \sum_{T \in \mathcal{S}} w(T)$$

and

$$\det L[\overline{\{j,n\}}, \overline{\{i,n\}}] = \sum_{F \in \mathcal{S}_v^{\{i,j\}}} w(F) \quad \text{for } 1 \leq i,j \leq n-1.$$

Therefore, (2.1) implies the following theorem.

THEOREM 2.3. *Let $\mathcal{G}$ be a connected weighted graph on vertices $1, 2, \ldots, n$. Let v be a vertex of $\mathcal{G}$. Then the bottleneck matrix $M = [m_{i,j}]$ of $\mathcal{G}$ based at vertex v satisfies*

$$m_{i,j} = \frac{\sum_{F \in \mathcal{S}_v^{\{i,j\}}} w(F)}{\sum_{T \in \mathcal{S}} w(T)}.$$

Theorem 2.3 has numerous consequences. The first, which is immediate, is a combinatorial formula for the bottleneck matrix of an unweighted graph.

COROLLARY 2.4. *Let $\mathcal{G}$ be an unweighted connected graph with n vertices. Then the bottleneck matrix, $M = [m_{i,j}]$, based at vertex v satisfies*

$$m_{i,j} = \frac{|\mathcal{S}_v^{\{i,j\}}|}{|\mathcal{S}|}.$$

Let $\mathcal{G}$ be an unweighted connected graph with vertices $1, 2, \ldots, n$. Let i and j be adjacent vertices. Adding the edge joining i and j to each forest in $\mathcal{S}_j^{\{i,i\}}$ establishes a correspondence between $\mathcal{S}_n^{\{i,i\}}$ and the spanning trees of $\mathcal{G}$ which contain the edge joining i and j. Hence, by Corollary 2.4, the (i,i)-entry of the bottleneck matrix of $\mathcal{G}$ based at vertex j is equal to the fraction of spanning trees of $\mathcal{G}$ which contain the edge joining i and j. An analogous result holds for weighted graphs and follows from Theorem 2.3.

Let T be a weighted tree with vertices $1, 2, \ldots, n$. Let i and j be (not necessarily distinct) vertices other than n of T. Each spanning forest F of T with exactly two components can be obtained from T by removing exactly one edge e. Thus, a spanning forest F of T is an $(\{i,j\}, n)$-spanning forest if and only if F is a spanning forest obtained from T by removing one of the edges e which lies on the path from i to n and on the path from j to n. Since $\frac{w(F)}{w(T)} = \frac{1}{w(e)}$, Theorem 2.3 implies the following.

COROLLARY 2.5. *Let T be a weighted tree with vertices 1, 2, ..., n. The bottleneck matrix, $M = [m_{i,j}]$, of $\mathcal{G}$ based at vertex v satisfies*

$$m_{i,j} = \sum_{e \in \mathcal{P}_{i,j}^v} \frac{1}{w(e)},$$

where $\mathcal{P}_{i,j}^v$ is the set of all edges e which lie on the path from i to v and on the path from j to v.

A direct proof of Corollary 2.5 is given in [8]. Note that if T is an unweighted tree, then by Corollary 2.5, $m_{i,j} = |\mathcal{P}_{i,j}^v|$. This was shown in [7].

Let $A = [a_{i,j}]$ be an $n \times n$ matrix. The matrix A is *combinatorially symmetric* provided $a_{i,j} = 0$ whenever $a_{j,i} = 0$. If A is combinatorially symmetric, then the *graph of* A is the graph $\mathcal{G}$ with vertices $1, 2, \ldots, n$ with an edge joining vertex i and vertex j if and only if $i \neq j$ and $a_{i,j} \neq 0$. If A is symmetric, then the *weighted graph* of A is the graph obtained from $\mathcal{G}$ by weighting the edge joining vertex i and j by $a_{i,j}$. The next result extends Corollary 2.5 to symmetric M-matrices whose graph is a tree.

COROLLARY 2.6. *Let A be a symmetric, singular M-matrix whose graph is a tree T with n vertices. Let x be a right null vector of A. Then the bottleneck matrix, $M = [m_{i,j}]$, of A based at v satisfies*

$$m_{i,j} \;=\; x_i x_j \sum_{e=\{k,\ell\} \in \mathcal{P}^v_{i,j}} \frac{1}{w(e) x_k x_\ell}.$$

Proof. Without loss of generality we take $v = n$. Let D be the $n \times n$ diagonal matrix whose diagonal entries are the entries of x. Consider the matrix $L = DAD$. It is easy to verify that L is the Laplacian matrix of the weighted tree obtained from T by weighting the edge e by $\overline{w}(e) = w(e) x_k x_\ell$, where k and ℓ are the vertices incident to e. Hence, by Corollary 2.5,

$$(L[\{n\}, \{n\}])^{-1} \;=\; \sum_{e \in \mathcal{P}^n_{i,j}} \frac{1}{\overline{w}(e)} \;=\; \sum_{e=\{k,\ell\} \in \mathcal{P}^n_{i,j}} \frac{1}{w(e) x_k x_\ell}.$$

The corollary now follows from the observation that

$$A[\{n\}, \{n\}]^{-1} \;=\; \hat{D} (L[\{n\}, \{n\}])^{-1} \hat{D},$$

where $\hat{D}$ is the diagonal matrix obtained from D by deleting row and column n. ∎

Corollary 2.6 can be extended to combinatorially symmetric singular M-matrices whose graph is a tree. First, it is shown that every such matrix is diagonally similar to a symmetric matrix. This implies that in studying the diagonal entries of the group inverse of a combinatorially symmetric singular M-matrix, one may assume that the matrix is in fact symmetric. The following lemma is, essentially, due to Parter and Youngs [10], though the proof below is different.

LEMMA 2.7 (see Parter and Youngs [10, Lemma 3]). *Let A be a combinatorially symmetric, singular M-matrix of order n whose graph is a tree T. Let $x = (x_1, x_2, \ldots, x_n)^T$ and $y = (y_1, y_2, \ldots, y_n)^T$ be right and left null vectors of A, respectively. Let $D = \operatorname{diag}(\frac{\sqrt{x_1}}{\sqrt{y_1}}, \frac{\sqrt{x_2}}{\sqrt{y_2}}, \ldots, \frac{\sqrt{x_n}}{\sqrt{y_n}})$. Then $D^{-1}AD$ is a symmetric, singular M-matrix whose graph is T.*

Proof. The proof is by induction on n. The result is clearly true if $n = 1$ or $n = 2$. Now assume that $n \geq 3$, and the result is true for any combinatorially symmetric, singular M-matrix of order $n - 1$ whose graph is a tree. Consider the $n \times n$ matrix $A = [a_{i,j}]$ and its graph T. Without loss of generality we may assume that vertex n is a pendant vertex in T and is adjacent to vertex $n - 1$. Since $Ax = 0$ and $y^T A = 0$,

$$\frac{y_{n-1}}{y_n} a_{n-1,n} \;=\; -a_{n,n} \;=\; \frac{x_{n-1}}{x_n} a_{n,n-1}.$$

Thus, it follows that the last row and column of $D^{-1}AD$ are transposes of each other. The matrix

$$\hat{A} \;=\; A[\overline{\{n\}}, \overline{\{n\}}] - \frac{x_n}{x_{n-1}} a_{n-1,n} E_{n-1,n-1},$$

where $E_{n-1,n-1}$ is the $n-1 \times n-1$ matrix with a 1 in position $(n-1, n-1)$ and 0 elsewhere, is a singular M-matrix whose graph is the tree obtained from T by deleting vertex n. In addition,

$$\hat{A} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \end{bmatrix} = 0 \text{ and } \begin{bmatrix} y_1 & y_2 & \cdots & y_{n-1} \end{bmatrix} \hat{A} = 0.$$

Thus, it follows by induction that $\hat{D}^{-1}\hat{A}\hat{D}$ is symmetric where

$$\hat{D} = \operatorname{diag}\left(\frac{\sqrt{x_1}}{\sqrt{y_1}}, \frac{\sqrt{x_2}}{\sqrt{y_2}}, \ldots, \frac{\sqrt{x_{n-1}}}{\sqrt{y_{n-1}}}\right).$$

Since $\hat{D}^{-1}A[\overline{\{n\}}, \overline{\{n\}}]\hat{D} = (D^{-1}AD)[\overline{\{n\}}, \overline{\{n\}}]$, we conclude that $D^{-1}AD$ is symmetric. □

COROLLARY 2.8. *Let A be an $n \times n$ combinatorially symmetric, singular M-matrix whose graph is a tree T and whose right and left null vectors are $x = (x_1, x_2, \ldots, x_n)^T$ and $y = (y_1, y_2, \ldots, y_n)^T$, respectively. Then the bottleneck matrix $M = [m_{i,j}]$ of A based at v satisfies*

$$m_{i,j} = x_j y_i \sum_{e=\{k,\ell\} \in \mathcal{P}^v_{i,j}} \frac{1}{a_{k\ell} x_\ell y_k}.$$

Proof. Without loss of generality we assume that $v = n$. Let $B = D^{-1}AD$, where $D = \operatorname{diag}(\frac{\sqrt{x_1}}{\sqrt{y_1}}, \frac{\sqrt{x_2}}{\sqrt{y_2}}, \ldots, \frac{\sqrt{x_n}}{\sqrt{y_n}})$. By Lemma 2.7, $B = [b_{i,j}]$ is a symmetric, singular M-matrix whose graph is T. It is easy to verify that

$$z = (\sqrt{x_1 y_1}, \sqrt{x_2 y_2}, \ldots, \sqrt{x_n y_n})^T$$

is a right null vector of B, and that the weight of an edge joining vertices k and ℓ equals $a_{k\ell}\frac{\sqrt{x_\ell y_k}}{\sqrt{x_k y_\ell}}$. Hence by Corollary 2.6, the (i,j)-entry of $(B[\overline{\{n\}}, \overline{\{n\}}])^{-1}$ equals

$$\sqrt{x_i y_i}\sqrt{x_j y_j} \sum_{e=\{k,\ell\} \in \mathcal{P}_{i,j}} \frac{1}{\left(a_{k\ell}\frac{\sqrt{x_\ell y_k}}{\sqrt{x_k y_\ell}}\sqrt{x_k y_k}\sqrt{x_\ell y_\ell}\right)},$$

which simplifies to

$$\sqrt{x_i x_j y_i y_j} \sum_{e=\{k,\ell\} \in \mathcal{P}_{i,j}} \frac{1}{a_{k,\ell} x_\ell y_k}.$$

Since $B[\overline{\{n\}}, \overline{\{n\}}] = \hat{D}^{-1}A[\overline{\{n\}}, \overline{\{n\}}]\hat{D}$, where $\hat{D} = D[\overline{\{n\}}, \overline{\{n\}}]$,

$$m_{i,j} = x_j y_i \sum_{e=\{k,\ell\} \in \mathcal{P}_{i,j}} \frac{1}{a_{k,\ell} x_\ell y_k}. \qquad \square$$

Next we derive a formula for the diagonal entries of Laplacians of trees. Let T be a tree. If e is an edge, then $T \setminus e$ denotes the graph obtained from T by removing

e. If i is a vertex of T, then we define $\beta_i(e)$ to be the set of vertices in the connected component of $T \setminus e$ which does not contain vertex i.

LEMMA 2.9. *Let $A = [a_{i,j}]$ be an $n \times n$ symmetric, singular M-matrix whose graph is a tree T and let $x = (x_1, x_2, \ldots, x_n)^T$ be a null vector of A. Then, for $v = 1, 2, \ldots, n$,*

$$A^{\#}_{v,v} = \frac{(x_v)^2}{(x^T x)^2} \sum_{e=\{k,\ell\} \in T} \frac{1}{w(e) x_k x_\ell} \left(\sum_{i \in \beta_v(e)} x_i^2 \right)^2 .$$

Proof. Without loss of generality we may assume that $v = n$. By Proposition 2.1,

$$A^{\#}_{n,n} = \frac{(x_n)^2}{(x^T x)^2} \hat{x}^T M \hat{x},$$

where $\hat{x}$ is the vector obtained from x by deleting its last row. Therefore, by Corollary 2.6,

$$\begin{aligned} A^{\#}_{n,n} &= \frac{(x_n)^2}{(x^T x)^2} \left(\sum_{1 \le i,j \le n-1} x_i^2 x_j^2 \sum_{e=\{k,\ell\} \in \mathcal{P}^n_{i,j}} \frac{1}{w(e) x_k x_\ell} \right) \\ &= \frac{(x_n)^2}{(x^T x)^2} \sum_{e \in T} \frac{1}{w(e) x_k x_\ell} \left(\sum_{\substack{1 \le i,j \le n-1 \\ e \in \mathcal{P}^n_{i,j}}} x_i^2 x_j^2 \right) \\ &= \frac{(x_n)^2}{(x^T x)^2} \sum_{e \in T} \frac{1}{w(e) x_k x_\ell} \left(\sum_{i \in \beta_n(e)} x_i^2 \right)^2 . \end{aligned}$$

The last equality follows from the fact that i and j are vertices such that $e \in \mathcal{P}^n_{i,j}$ if and only if both i and j belong to $\beta_n(e)$. ∎

We have the following immediate consequence for the Laplacian matrix of a weighted tree. The analogous result for the Laplacian matrix of an unweighted tree is contained in Theorem 3.3 of [7].

COROLLARY 2.10. *If L is the Laplacian matrix of a weighted tree T with n vertices, then*

$$L^{\#}_{v,v} = \frac{1}{n^2} \sum_{e \in T} \frac{|\beta_v(e)|^2}{w(e)} \tag{2.2}$$

for $v = 1, 2, \ldots, n$.

The analogous result for nonsymmetric matrices follows by a similar argument.

COROLLARY 2.11. *If A is an $n \times n$ combinatorially symmetric, singular M-matrix whose graph is a tree and x and y are right and left null vectors, respectively, of A, then*

$$A^{\#}_{v,v} = \frac{x_v y_v}{(y^T x)^2} \sum_{e=\{k,\ell\} \in T} \frac{1}{a_{k,\ell} x_\ell y_k} \left(\sum_{i \in \beta_v(e)} x_i y_i \right)^2$$

for $v = 1, 2, \ldots, n$.

We now give a formula for the difference between certain diagonal entries.

LEMMA 2.12. *Let* $A = [a_{ij}]$ *be an* $n \times n$ *symmetric, singular* M*-matrix whose graph is a tree* T *and let* $x = (x_1, x_2, \ldots, x_n)^T$ *be the null vector of* A*. Assume that vertices* i *and* j *are joined by an edge* e*. Then*

$$\frac{1}{x_i^2}A^{\#}_{i,i} - \frac{1}{x_j^2}A^{\#}_{j,j} = \frac{1}{(x^Tx)^2}\frac{1}{w(e)x_ix_j}\left[\left(\sum_{k\in\beta_i(e)} x_k^2\right)^2 - \left(\sum_{k\in\beta_j(e)} x_k^2\right)^2\right].$$

Proof. Let f be an edge of T with $f \neq e$. Then $\beta_j(f) = \beta_i(f)$. The result now follows from Lemma 2.9. □

For weighted trees we have the following lemma.

LEMMA 2.13. *Let* T *be a weighted tree on* n *vertices with Laplacian matrix* L*. Suppose that* i *and* j *are vertices of* T *joined by the edge* e*. Then*

$$L^{\#}_{i,i} - L^{\#}_{j,j} = \frac{1}{nw(e)}(|\beta_i(e)| - |\beta_j(e)|). \tag{2.3}$$

In particular, $L^{\#}_{i,i} > L^{\#}_{j,j}$ *if and only if* $|\beta_i(e)| > |\beta_j(e)|$.

Proof. Since $\mathbf{1}$ is a null vector of L, and $|\beta_i(e)| + |\beta_j(e)| = n$, the result follows from Lemma 2.12. □

Finally, in the next section we shall also require a formula for the off-diagonal entries of the Laplacian of a weighted tree.

LEMMA 2.14. *Let* T *be a weighted tree on* n *vertices with Laplacian matrix* L*. Then, for* $i, j = 1, \ldots, n$ *with* $i \neq j$*,*

$$L^{\#}_{i,j} = \frac{1}{n^2}\sum_{e\in T}\frac{|\beta_j(e)|^2}{w(e)} - \frac{1}{n}\sum_{e\in\mathcal{P}_{i,j}}\frac{|\beta_j(e)|}{w(e)}. \tag{2.4}$$

Proof. Without loss of generality we can assume that $j = n$ and $i = 1, \ldots, n-1$. Then using the symmetry of L, it follows from Proposition 2.2 and Corollary 2.6 that

$$\begin{aligned} L^{\#}_{n,i} &= \frac{1}{n^2}\mathbf{1}^TM\mathbf{1} - \frac{1}{n}\sum_{k=1}^{n-1}\sum_{e\in\mathcal{P}^n_{i,k}}\frac{1}{w(e)} \\ &= \frac{1}{n^2}\sum_{e\in T}\frac{|\beta_j(e)|^2}{w(e)} - \frac{1}{n}\sum_{e\in\mathcal{P}_{i,n}}\frac{|\beta_i(e)|}{w(e)}. \end{aligned}$$ □

3. Inverse weighted distances. We begin with the following auxilliary lemma.

LEMMA 3.1. *Let* T *be a weighted tree on* n *vertices. Let* v_0 *and* v_l *be vertices in* T *with* $v_1, v_2, \ldots, v_{l-1}$ *as intermediate vertices on the path* α *which joins* v_0 *to* v_l*. For* $1 \leq i \leq l$*, let* e_i *be the edge between* v_{i-1} *and* v_i *having weight* θ_i*. For* $0 \leq i \leq l$*, let* t_i *be the number of vertices, including* v_i *whose shortest path to* α *has terminal vertex* v_i*. Then*

$$\tilde{d}_{v_0} - \tilde{d}_{v_l} = \sum_{i=0}^{l} t_i\left(\sum_{m=1}^{i}\frac{1}{\theta_m} - \sum_{m=i+1}^{l}\frac{1}{\theta_m}\right). \tag{3.1}$$

Proof. For any one of the t_i vertices u of T whose shortest path to α ends at v_i, we have that

$$\tilde{d}(u,v_0) - \tilde{d}(u,v_l) \;=\; \tilde{d}(v_i,v_0) - \tilde{d}(v_i,v_l).$$

Hence we find that

$$\tilde{d}_{v_0} - \tilde{d}_{v_l} \;=\; \sum_{i=1}^{l} t_i \left[\tilde{d}(v_i,v_0) - \tilde{d}(v_i,v_l)\right].$$

The result now follows on observing that

$$\tilde{d}(v_i,v_0) - \tilde{d}(v_i,v_l) \;=\; \left(\sum_{m=1}^{i} \frac{1}{\theta_m} - \sum_{m=i+1}^{l} \frac{1}{\theta_m}\right). \qquad \square$$

Lemma 3.1 leads us to the following theorem.

THEOREM 3.2. *Under the assumptions and notations of Lemma* 3.1 *we have that*

$$n\left(L^{\#}_{v_0,v_0} - L^{\#}_{v_l,v_l}\right) \;=\; \sum_{i=0}^{l} t_i \left(\sum_{m=1}^{i} \frac{1}{\theta_m} - \sum_{m=i+1}^{l} \frac{1}{\theta_m}\right). \tag{3.2}$$

Proof. From (2.2) we have that

$$n^2 L^{\#}_{v_0,v_0} \;=\; \sum_{e\in T} \frac{\beta^2_{v_0}(e)}{w(e)}$$

and that

$$n^2 L^{\#}_{v_l,v_l} \;=\; \sum_{e\in T} \frac{|\beta^2_{v_l}(e)|}{w(e)}.$$

Now if $e \notin \alpha$, then $|\beta_{v_0}(e)| = |\beta_{v_l}(e)|$, while if $e = e_m$ for some $m = 1, 2, \ldots, \ell$, then

$$|\beta_{v_0}(e_m)| \;=\; t_m + \cdots + t_l$$

and

$$|\beta_{v_l}(e_m)| \;=\; t_0 + \cdots + t_{m-1}.$$

Hence,

$$\begin{aligned} n^2\left(L^{\#}_{v_0,v_0} - L^{\#}_{v_l,v_l}\right) &= \textstyle\sum_{m=1}^{l} \left(|\beta_{v_0}(e_m)| + |\beta_{v_l}(e_m)|\right) \frac{|\beta_{v_0}(e_m)| - |\beta_{v_l}(e_m)|}{w(e_m)} \\ &= n \textstyle\sum_{m=1}^{l} \frac{-t_0 - \cdots - t_{m-1} + t_m + \cdots + t_l}{\theta_m}. \end{aligned}$$

In the above sum, we collect terms in each t_i to find that

$$n^2\left(L^{\#}_{v_0,v_0} - L^{\#}_{v_l,v_l}\right) \;=\; n\sum_{i=0}^{l} t_i \left(\sum_{m=1}^{i} \frac{1}{\theta_m} - \sum_{m=i+1}^{l} \frac{1}{\theta_m}\right). \qquad \square$$

COROLLARY 3.3. *Let T be a weighted tree on n vertices and L be its Laplacian. Then for some constant c,*

$$\left(L^{\#}_{1,1},\ldots,L^{\#}_{n,n}\right) \;=\; \frac{1}{n}(\tilde{d}_1,\ldots,\tilde{d}_n)+c\mathbf{1}. \tag{3.3}$$

Proof. Let Z be the simple cycle permutation matrix sending $1\to n$ and $i\to i-1,\ i=2,\ldots,n$. It follows from Lemma 3.1 and Theorem 3.2 that

$$(I-Z)\left[\left(L^{\#}_{1,1},\ldots,L^{\#}_{n,n}\right)^T-\frac{1}{n}\left(\tilde{d}_1,\ldots,\tilde{d}_n\right)^T\right] \;=\; 0$$

from which the conclusion easily follows. ∎

This corollary shows that if $i_1,\ldots,i_n$ are indices of entries in $(L^{\#}_{1,1},\ldots,L^{\#}_{n,n})^T$ such that

$$L^{\#}_{i_1,i_1} \;\geq\; L^{\#}_{i_2,i_2} \;\geq\cdots\geq\; L^{\#}_{i_n,i_n},$$

then

$$\tilde{d}_{i_1,i_1} \;\geq\; \tilde{d}_{i_2,i_2} \;\geq\cdots\geq \tilde{d}_{i_n,i_n}.$$

In particular we have the following conclusion.

COROLLARY 3.4. *Let T be a weighted tree on n vertices and L its Laplacian. Then a diagonal entry in $L^{\#}$ occurs at an index which corresponds to an index of a pendant vertex of T whose inverse status is maximal and a minimal diagonal entry in $L^{\#}$ occurs at an index which corresponds to vertex of T which is a centroid.*

Proof. To prove the first part of the corollary, suppose that k is not a pendant vertex, say with k adjacent to m, with edge e_1 between them, and k is adjacent to l with edge e_2 between them. If $L^{\#}_{k,k}-L^{\#}_{m,m}\geq 0$, then, by (2.3),

$$|\beta_k(e_1)| \;\geq\; |\beta_m(e_1)|.$$

But as $|\beta_k(e_1)|+|\beta_m(e_1)|=n$, we find that

$$|\beta_k(e_1)| \;\geq\; \frac{n}{2}.$$

Now, as $k\in\beta_k(e_1)\cup\beta_k(e_2)$, we find that

$$|\beta_k(e_2)| \;\leq\; \frac{n-2}{n},$$

so that $L^{\#}_{k,k}-L^{\#}_{l,l}<0$. Hence $L^{\#}_{k,k}$ cannot be the maximal diagonal entry in $L^{\#}$. It now follows that the maximal diagonal entry must occur at a pendant vertex.

To see that the minimal inverse status number occurs at a centroid, note first that on the one hand vertex l is centroid if and only if

$$|\beta_l(e)| \;\leq\; \frac{n}{2}$$

for all edges e incident with l. On the other hand, because for any adjacent vertices i and j, $|\beta_j(e)|+|\beta_i(e)|=n$, where e is the adjacent edge between i and j, in order for vertex j to satisfy that

$$L^{\#}_{j,j} \;=\; \min_{1\leq i\leq n} L^{\#}_{i,i},$$

it is necessary, again by (2.3), that $|\beta_j(e)| \le n/2$ for all edges incident with j. Our proof is now complete. □

For each edge $e \in T$, $T \setminus \{e\}$ has two connected components, J_e has j_e vertices and $N \setminus J_e$ has $n - j_e$ vertices.

THEOREM 3.5.

$$\sum_{i=1}^{n} \tilde{d}_i \;=\; 2n \sum_{i=1}^{n} L^{\#}_{i,i} \;=\; \sum_{e \in T} \frac{2j_e(n-j_e)}{w(e)}.$$

Proof. From the definition of a inverse status number in (1.2) we have that

$$\sum_{i=1}^{n} \tilde{d}_i \;=\; \sum_{v \in T} \sum_{u \in T} \sum_{e \in \mathcal{P}_{u,v}} \frac{1}{w(e)}.$$

For each edge e, there are $2j_e(n-j_e)$ unordered pairs of vertices u and v such that e is on the path between them. Hence, each edge e contributes $2j_e(n-j_e)/w(e)$ to the above sum so that

$$\sum_{i=1}^{n} \tilde{d}_i \;=\; \sum_{e \in T} \frac{2j_e(n-j_e)}{w(e)}.$$

Now from (2.2),

$$2n \sum_{i=1}^{n} L^{\#}_{i,i} \;=\; \frac{2}{n} \sum_{v \in T} \sum_{e \in T} \frac{|\beta_v(e)|^2}{w(e)} \;=\; \frac{2}{n} \sum_{e \in T} \sum_{v \in T} \frac{|\beta_v(e)|^2}{w(e)}.$$

If $v \in J_e$, then

$$|\beta_v(e)|^2 \;=\; (n - j_e)^2,$$

while if $v \in N \setminus J_e$, then

$$|\beta_v(e)|^2 \;=\; j_e^2.$$

Consequently,

$$\begin{aligned} \textstyle\sum_{e \in T} \sum_{v \in T} \frac{|\beta_v(e)|^2}{w(e)} &= \textstyle\sum_{e \in T} \frac{1}{w(e)} \left(\sum_{v \in J_e} |\beta_v(e)|^2 + \sum_{v \in N \setminus J_e} |\beta_v(e)|^2 \right) \\ &= \textstyle\sum_{e \in T} \frac{1}{w(e)} \left((n - j_e)^2 j_e + j_e^2 (n - j_e) \right) \\ &= \textstyle\sum_{e \in T} \frac{n j_e (n - j_e)}{w(e)}. \end{aligned}$$

The result now follows. □

We can now give a precise value to the constant c of (3.3).

COROLLARY 3.6.

$$\left(L^{\#}_{1,1}, \ldots, L^{\#}_{n,n} \right) = \frac{1}{n} (\tilde{d}_1, \ldots, \tilde{d}_n) - \frac{1}{n^2} \sum_{e \in T} \frac{j_e(n-j_e)}{w(e)} \mathbf{1} \;=\; \frac{1}{n} (\tilde{d}_1, \ldots, \tilde{d}_n) - \frac{1}{2n^2} \sum_{i=1}^{n} \tilde{d}_i \mathbf{1}.$$

Proof.

$$c\mathbf{1} \;=\; \left(L^{\#}_{1,1},\ldots,L^{\#}_{n,n}\right) - \frac{1}{n}(\tilde{d}_1,\ldots,\tilde{d}_n),$$

so that

$$nc \;=\; \frac{1}{n}\sum_{e\in T}\frac{j_e(n-j_e)}{w(e)} - \frac{2}{n}\sum_{e\in T}\frac{j_e(n-j_e)}{w(e)}.$$

Solving for c now yields the result. ∎

Remark. We comment that as the nonzero eigenvalues of the group inverse of the Laplacian are the reciprocals of the nonzero eigenvalues of the Laplacian, Theorem 3.5 generalizes an equality for the *Wiener index* of an unweighted tree (see Merris [9, Theorem 5.5] and references cited therein), namely, that

$$\sum_{i=1}^{n}\tilde{d}_i \;=\; n\sum_{\lambda\in\sigma(L)\setminus\{0\}}\frac{1}{\lambda},$$

where $\sigma(\cdot)$ denotes the spectrum of a matrix.

We next develop formulas, in terms of distances, for the off-diagonal entries in $L^{\#}$.

THEOREM 3.7. *For $i\neq j$, $1\leq i,k\leq n$,*

$$L^{\#}_{i,n} \;=\; \frac{\tilde{d}_i+\tilde{d}_k}{2n} - \frac{1}{2}\tilde{d}_{i,k} - \frac{1}{2n^2}\sum_{j=1}^{n}\tilde{d}_j. \tag{3.4}$$

Proof. Without loss of generality we can take $k=n$ and $i=1,\ldots,n-1$. First we claim that

$$L^{\#}_{i,i} - L^{\#}_{n,n} \;=\; \tilde{d}_{i,n} - \frac{2}{n}\sum_{e\in\mathcal{P}_{i,n}}\frac{\beta_n(e)}{w(e)}.$$

To see this note that if $e\notin\mathcal{P}_{i,n}$, then its contribution to $L^{\#}_{i,i}$ is the same as that to $L^{\#}_{n,n}$. On the other hand, if $e\in\mathcal{P}_{i,n}$, the contribution of e to $L^{\#}_{n,n}$ is

$$\frac{1}{n}\frac{\beta_n(e)}{w(e)},$$

while its contribution to $L^{\#}_{i,i}$ is

$$\frac{1}{n}\frac{n-\beta_n(e)}{w(e)}.$$

Thus

$$L^{\#}_{i,i} - L^{\#}_{n,n} \;=\; \frac{2}{n}\sum_{e\in\mathcal{P}_{i,n}}\frac{n-\beta_n(e)}{w(e)} \;=\; \tilde{d}_{i,n} - \frac{2}{n}\sum_{e\in\mathcal{P}_{i,n}}\frac{\beta_n(e)}{w(e)},$$

as desired. Hence, for $1\leq j\leq n$, we have that

$$L^{\#}_{j,j} \;=\; \tilde{d}_{j,n} - \frac{2}{n}\sum_{f\in\mathcal{P}_{i,n}}\frac{\beta_n(f)}{w(f)},$$

and this yields that

$$\begin{aligned} L^{\#}_{i,i} - L^{\#}_{j,j} &= \tilde{d}(i,n) - \tilde{d}(j,n) + 2\left[\frac{1}{n}\sum_{f\in\mathcal{P}_{j,n}}\frac{\beta_n(f)}{w(f)} - \frac{1}{n}\sum_{f\in\mathcal{P}_{i,n}}\frac{\beta_n(e)}{w(e)}\right] \\ &= \tilde{d}(i,n) - \tilde{d}(j,n) + 2\left(L^{\#}_{i,n} - L^{\#}_{n,n}\right). \end{aligned}$$

It now follows that for some constant α,

$$2L^{\#}_{i,n} = L^{\#}_{i,i} - \tilde{d}(i,n) + \alpha, \quad 1 \le i \le n-1.$$

To find α, note that

$$\begin{aligned} (n-1)\alpha &= 2\sum_{i=1}^{n-1} L^{\#}_{i,n} - 2\sum_{j=1}^{n-1} L^{\#}_{j,n} + \sum_{i=1}^{n-1}\tilde{d}(i,n) \\ &= -2L^{\#}_{n,n} - \sum_{i=1}^{n-1} L^{\#}_{i,i} + \tilde{d}_n = -L^{\#}_{n,n} - \sum_{i=1}^{n} L^{\#}_{i,i} + \tilde{d}_n \\ &= -\frac{1}{n}\tilde{d}_n + \frac{1}{2n^2}\sum_{j=1}^{n}\tilde{d}_j - \frac{1}{2n}\sum_{j=1}^{n}\tilde{d}_j + \tilde{d}_n \\ &= \frac{n-1}{n}\tilde{d}_n - \frac{n-1}{2n^2}\sum_{j=1}^{n}\tilde{d}_j. \end{aligned}$$

Thus we have that

$$\alpha = \frac{1}{n}\tilde{d}_n - \frac{1}{2n^2}\sum_{j=1}^{n}\tilde{d}_j,$$

which yields

$$\begin{aligned} L^{\#}_{i,n} &= \frac{1}{2}\left(L^{\#}_{i,i} - \tilde{d}(i,n) + \frac{1}{n}\tilde{d}_n - \frac{1}{2n^2}\sum_{j=1}^{n}\tilde{d}_j\right) \\ &= \frac{1}{2}\left[\frac{1}{n}\left(\tilde{d}_i + \tilde{d}_n\right) - \tilde{d}(i,n) - \frac{1}{n^2}\sum_{j=1}^{n}\tilde{d}_j\right], \end{aligned}$$

as desired. □

In Deutsch and Neumann [5], the following problem was posed: characterize the set of all $n \times n$ irreducible singular M-matrices whose group inverse is also an M-matrix. In the circumstances of this paper we have the following result.

COROLLARY 3.8. *Let L be the Laplacian of a weighted tree on n vertices. Then $L^{\#}$ is an M-matrix if and only if for every pair of adjacent vertices i and j we have that*

$$\tilde{d}_i + \tilde{d}_j \le \frac{n}{w(e_{i,j})} + \frac{1}{n}\sum_{k=1}^{n}\tilde{d}_k.$$

Proof. For vertices a and b with $1 \leq a, b \leq n-1$, we have from (2.4) that

$$L^{\#}_{a,n} - L^{\#}_{b,n} \;=\; \frac{1}{n}\left(\sum_{f\in\mathcal{P}_{b,n}} \frac{\beta_n(f)}{w(f)} - \sum_{e\in\mathcal{P}_{a,n}} \frac{\beta_n(e)}{w(e)}\right),$$

so we see that, in the nth row of $L^{\#}$, the entries are decreasing along paths away from n. Hence $L^{\#}$ is an M-matrix if and only if for any adjacent vertices i and j we have that $L^{\#}_{i,j} \leq 0$. Imposing this condition in (3.4) now yields the result. ∎

For unweighted trees we can determine precisely which trees admit a Laplacian whose group inverse is an M-matrix. For this purpose we require the following lemma.

PROPOSITION 3.9. *Let T be an unweighted tree on n vertices. Then the maximum of $\tilde{d}_i + \tilde{d}_j$ over all pairs of adjacent vertices i and j occurs for the adjacent vertices of some pendant edge.*

Proof. Let i and j be adjacent vertices with $\tilde{d}_i + \tilde{d}_j$ maximal. Assume that neither i nor j is a pendant vertex. Then there exist vertices $h \neq j$ and $k \neq i$ such that h is adjacent to i and k is adjacent to j. Let α, β, γ, and δ, respectively, be the set of vertices v for which the path from v to the path h—i—j—k ends at h, i, j, and k, respectively.

Note that the distance from a vertex in γ to either i or k is equal, the distance from a vertex in δ to i is 2 more than its distance to k, and the distance from a vertex in $\alpha \cup \beta$ to k is 2 more than its distance to i. Thus,

$$\tilde{d}_i - \tilde{d}_k \;=\; 2(|\delta| - |\alpha| - |\beta|). \tag{3.5}$$

Similarly,

$$\tilde{d}_j - \tilde{d}_h \;=\; 2(|\alpha| - |\gamma| - |\delta|). \tag{3.6}$$

Since $\tilde{d}_i + \tilde{d}_j$ is maximal, it follows that $\tilde{d}_i - \tilde{d}_k \geq 0$ and $\tilde{d}_j - \tilde{d}_h \geq 0$. Adding equations (3.5) and (3.6) yields

$$0 \;\leq\; -2(|\beta| + |\gamma|).$$

This implies that β and γ are empty. However, this contradicts the fact that $i \in \beta$ and $j \in \gamma$. We conclude that either i or j is a pendant vertex. ∎

We can now show that for unweighted trees, the only trees whose Laplacian has a group inverse which is an M-matrix are the stars of all orders.

THEOREM 3.10. *Let L be the Laplacian of a tree T with n vertices. Then $L^{\#}$ is an M-matrix if and only if T is a star.*

Proof. By Corollary 3.8,

$$d_i + d_j \;\leq\; n + \frac{1}{n}\sum_{k=1}^{n} \tilde{d}_k \tag{3.7}$$

for each pair of adjacent vertices i and j. Let i be a pendant vertex and j the vertex adjacent to i, and let e be the edge joining i and j. By counting the contributions of each edge of T to $\tilde{d}_i + \tilde{d}_j$ and to $\sum_{k=1}^{n} \tilde{d}_k$ we see that

$$\tilde{d}_i + \tilde{d}_j \;=\; n + 2\sum_{f\neq e} |\beta_i(f)|$$

and

$$\sum_{k=1}^{n} \tilde{d}_k \;=\; 2(n-1) + 2\sum_{f \neq e} |\beta_i(f)|(n - |\beta_i(f)|).$$

Substituting into (3.7) and simplifying we obtain that

$$\sum_{f \neq e} \beta_i(f)^2 \;\leq\; n-1.$$

Since $\beta_i(f) \geq 1$ for $f \neq e$, we conclude that $\beta_i(f) = 1$ for $f \neq e$. It now follows that T is a star with center j. □

We note that in Chen, Kirkland, and Neumann [4, Corollary 5.4] it is shown that if $n \geq 5$, then no weighted path of order n can yield a Laplacian whose group inverse is an M-matrix. The following example, taken from [4], exhibits a weighted path of order 4 whose group inverse is an M-matrix. Let

$$L \;=\; \begin{bmatrix} 0.2 & -0.2 & 0 & 0 \\ -0.2 & 0.6 & -0.4 & 0 \\ 0 & -0.4 & 0.6 & -0.2 \\ 0 & 0 & -0.2 & 0.2 \end{bmatrix}.$$

Then

$$L^{\#} \;=\; \begin{bmatrix} 3.75 & 0.000 & -1.25 & -2.5 \\ 0.000 & 1.250 & 0.000 & -1.250 \\ -1.25 & 0.000 & 1.25 & 0.000 \\ -2.50 & -1.25 & 0.000 & 3.75 \end{bmatrix}.$$

REFERENCES

[1] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, PA, 1994.

[2] S. L. Campbell and C. D. Meyer, Jr., *Generalized Inverses of Linear Transformations*, Dover Publications, New York, 1991.

[3] S. Chaiken, *A combinatorial proof of the all minors matrix tree theorem*, SIAM J. Alg. Disc. Meth., 3 (1982), pp. 319–329.

[4] Y. Chen, S. J. Kirkland, and M. Neumann, *Group generalized inverses of M-matrices associated with periodic and nonperiodic Jacobi matrices*, Linear and Multilinear Algebra, 39 (1995), pp. 325–340.

[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] F. Harary, *Status and contrastatus*, Sociometry, 22 (1959), pp. 23–43.

[7] S. J. Kirkland, M. Neumann, and B. Shader,.*Bounds on the subdominant eigenvalue involving group inverses with applications to graphs*, Czechoslovakia Math. J., to appear.

[8] S. J. Kirkland, M. Neumann, and B. Shader, *Characteristic vertices of weighted trees via perron values*, Linrear and Multilinear Algebra, 40 (1996), pp. 311–325.

[9] R. Merris, *Laplacian matrices and graphs: A survey*, Linear Algebra Appl., 197, 198 (1994), pp. 143–176.

[10] S. Parter and J. W. T. Youngs, *The symmetrization of matrices by diagonal matrices*, J. Math. Anal. Appl., 4 (1962), pp. 102–110.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 4, pp. 842–860, October 1997

004

ON THE BEHAVIOR OF A SEQUENCE DEFINED BY A PERIODIC RECURSIVE RELATION*

TIN-YAU TAM†

In memory of my mother-in-law, Sung-Chu Siu, who passed away on November 24, 1995. She was 76.

Abstract. The sequence $\{x_n\}$ defined by a periodic recursive relation of period r is examined. A necessary and sufficient condition is given for the subsequences $\{x_{nr+k}\}$, $k = 1, \ldots, r$, to be arithmetic progressions. Various generalizations are also considered.

Key words. periodic sequence, arithmetic progression, eigenvalue

AMS subject classifications. 11B37, 15A18

PII. S0895479896301972

1. Introduction.

Let λ be a real number. Let us consider the following sequence $\{x_n\}$ defined by the relation:

$$\begin{aligned} x_2 &= \lambda x_1 \\ x_1 + x_3 &= \lambda x_2 \\ &\vdots \\ x_{r-1} + x_{r+1} &= \lambda x_r \\ &\vdots \end{aligned}$$

where $x_1 \neq 0$. For example, if $\lambda = 0$, then $\{x_n\} = \{1, 0, -1, 0, 1, 0, -1, 0, \ldots\}$ by assigning $x_1 = 1$. What value(s) of λ will make the sequence $\{x_n\}$ in arithmetic progression? The answer is simple: $\lambda = 2$ and $x_n = nx_1$. Subtracting the second equation from the third equation, we have

$$(x_2 - x_1) + (x_4 - x_3) = \lambda(x_3 - x_2).$$

So if $x_{n+1} = x_n + d$ for $n = 1, 2, \ldots$, then $2d = \lambda d$.

If $d \neq 0$, then $\lambda = 2$. Adding the first $(n-1)$ equations yields $x_n - x_{n-1} = x_1$, $n = 1, 2, \ldots$, i.e., $d = x_1$. So $x_n = nx_1$.

If $d = 0$, then $\lambda = 1$ by the first equation. But then $x_1 = x_2 \neq 0$ and $x_3 = 0$. So the possibility $\lambda = 1$ is rejected.

Now we want to address the following generalized problem. Let λ be a real number and let $s_1, \ldots, s_r$ be r given real numbers. Then define the sequence $\{x_n\}$ by

*Received by the editors April 12, 1996; accepted for publication (in revised form) by R. Brualdi September 13, 1996.

http://www.siam.org/journals/simax/18-4/30197.html

†Department of Mathematics, Auburn University, Auburn, AL 36849-5310 (tamtiny@mail.auburn.edu).

the following periodic recursive relation:

$$
\begin{aligned}
s_1x_2 &= \lambda x_1\\
s_1x_1 + s_2x_3 &= \lambda x_2\\
&\vdots\\
s_{r-1}x_{r-1} + s_rx_{r+1} &= \lambda x_r\\
s_rx_r + s_1x_{r+2} &= \lambda x_{r+1}\\
s_1x_{r+1} + s_2x_{r+3} &= \lambda x_{r+2}\\
&\vdots\\
s_{r-1}x_{2r-1} + s_rx_{2r+1} &= \lambda x_{2r}\\
&\vdots
\end{aligned}
\tag{1}
$$

The example that we discussed is corresponding to $r = 1$ and $s_1 = 1$. The relation (1) can be rewritten in matrix version

$$Sx = \lambda x,$$

where $x = (x_1, x_2, \ldots)^T$ and S is the infinite tridiagonal matrix

$$
S = \begin{pmatrix}
0 & s_1 & & & & & & \\
s_1 & 0 & s_2 & & & & & \\
 & s_2 & \ddots & \ddots & & & & \\
 & & \ddots & 0 & s_r & & & \\
 & & & s_r & 0 & s_1 & & \\
 & & & & s_1 & 0 & s_2 & \\
 & & & & & s_2 & 0 & \ddots \\
 & & & & & & \ddots & \ddots
\end{pmatrix}.
\tag{2}
$$

If $r \geq 3$, denote by S' the $r \times r$ matrix

$$
S' = \begin{pmatrix}
0 & s_1 & & & & & s_r \\
s_1 & 0 & s_2 & & & & \\
 & s_2 & \ddots & \ddots & & & \\
 & & \ddots & & s_{r-2} & & \\
 & & & s_{r-2} & 0 & s_{r-1} & \\
s_r & & & & s_{r-1} & 0 &
\end{pmatrix}.
$$

Since S' is a real symmetric matrix, the eigenvalues of S' are real.

The relation (1) was investigated in [5] in order to solve a conjecture of Ridge [4] on the numerical range of a periodic weighted shift. Indeed it was shown in [5] that the Perron root of the nonnegative irreducible symmetric matrix S' is not an eigenvalue of the operator S although it is an approximate eigenvalue of S, where $s_1, \ldots, s_r > 0$ and $r \geq 3$.

When $r \geq 3$ and if one of the s's is zero, say, $s_r = 0$, then the vectors $(x_1, \ldots, x_r)^T$, $(x_{r+1}, \ldots, x_{2r})^T, \ldots$ are in the null space of $\lambda I - S'$.

When $r = 2$ and (i) if $s_1 = 0$ and $s_2 \neq 0$, then $\lambda = 0$ and $x_n = 0$ for $n = 2, 3, \ldots$, (ii) if $s_2 = 0$ and $s_1 \neq 0$, then $s_1 x_2 = \lambda x_1$ and $s_1 x_1 = \lambda x_2$. So $s_1^2 = \lambda^2$, i.e., $\lambda = \pm s_1$ and hence $x_{2n} = \pm x_{2n-1}$ for all positive integers n.

In order to avoid triviality, we assume that $x_1 \neq 0$ and $s_i \neq 0$ for all $i = 1, \ldots, r$ in (1).

We adopt the following notation. If B is an $n \times n$ matrix, denote by $B_{i,j}$, $1 \leq i < j \leq n$, the principal submatrix of B locating on the rows and the columns indexed by $i, i+1, \ldots, j$. We set $B_j \equiv B_{1,j}$. We denote by $\varphi_{i,j}(\lambda)$ the characteristic polynomial of $B_{i,j}$ and set $\varphi_j(\lambda) \equiv \varphi_{1,j}(\lambda)$.

When A is a Hermitian matrix, the algebraic multiplicity and the geometric multiplicity of an eigenvalue of A are the same. We simply call it multiplicity.

THEOREM 1. *Let $s_1, \ldots, s_r$ be r given nonzero real numbers. Let $\{x_n\}$ be the sequence defined by (1) with $x_1 \neq 0$. Then the subsequences $\{x_{nr+k}\}$, $k = 1, \ldots, r$, are in arithmetic progression; i.e., $x_{nr+k} = x_{(n-1)r+k} + d_k$, $k = 1, \ldots, r$, $n = 1, 2, \ldots$, if and only if*

(I) *$\lambda = 2s_1$ when $r = 1$,*

(II) *$\lambda = \pm(s_1 + s_2)$ when $r = 2$,*

(III) *λ is an eigenvalue of S' when $r \geq 3$.*

If (I) *happens, then $x_n = nx_1$, $n = 1, 2, \ldots$.*

If (II) *happens, then $x_{2n} = nx_2 = \pm\frac{n(s_1+s_2)}{s_1}x_1$ and $x_{2n+1} = x_1 \pm nx_2 = (1 + \frac{n(s_1+s_2)}{s_1})x_1$, $n = 1, 2, \ldots$.*

If (III) *happens and*

(a) *if λ is a simple eigenvalue of S', then the vector $d = (d_1, \ldots, d_r)^T$ is in the null space of $S' - \lambda I$ with $d_r = x_r$. In particular, $\{x_{nr}\} = \{nx_r\}$. More precisely, if z is a normalized eigenvector of S' corresponding to λ, and*

(i) *if $z_r = 0$, then $d = 0$, i.e., $x_{nr+k} = x_k$, $k = 1, \ldots, r-1$, and $x_{nr} = 0$ for all positive integers n. In other words, $\{x_n\}$ is a periodic sequence.*

(ii) *If $z_r \neq 0$, then*

$$d = \frac{\varphi'(\lambda) z_r}{s_1 \cdots s_{r-1}} z,$$

where $\varphi(\lambda)$ is the characteristic polynomial of S'. Hence d is an eigenvector of S' with $d_r = x_r$. Moreover, if $\lambda\,(> 0)$ is the largest eigenvalue of S', then $d_r > 0$.

(b) *If λ is an eigenvalue of S' of multiplicity 2, then $d = 0$; i.e., $x_{nr+k} = x_k$, $k = 1, \ldots, r-1$, and $x_{nr} = 0$ for all positive integers n. In other words, $\{x_n\}$ is a periodic sequence.*

There are no eigenvalues of S' with multiplicity greater than 2.

COROLLARY 1. *Let $r \geq 3$ and let $\varphi_j(\lambda)$ be the characteristic polynomial of S_j, $j = 1, 2, \ldots$.*

(a) *If λ is a simple eigenvalue of S' and if z is a normalized eigenvector of S' corresponding to λ such that $z_r \neq 0$, then*

$$\varphi_{nr+k-1}(\lambda) = s_1 \cdots s_r \varphi_{k-1}(\lambda) + n\frac{(s_1 \cdots s_r)^n \varphi'(\lambda) z_r z_k}{s_k \ldots s_{r-1}},$$

where $\varphi_0(\lambda) \equiv 1$, $k = 1, \ldots, r-1$, $n = 1, 2, \ldots$, and

$$\varphi_{nr-1}(\lambda) = n(s_1 \cdots s_r)^{n-1} \varphi_{r-1}(\lambda), \qquad n = 1, 2, \ldots .$$

(b) *If λ is a simple eigenvalue of S' and $z_r = 0$ or if λ is an eigenvalue of multiplicity 2, then*

$$\varphi_{nr+k-1}(\lambda) = (s_1 \cdots s_r)^n \varphi_{k-1}(\lambda), \qquad k = 1, \dots, r-1, \quad n = 1, 2, \dots,$$

where $\varphi_0(\lambda) \equiv 1$, and

$$\varphi_{nr-1}(\lambda) = 0, \qquad n = 1, 2, \dots.$$

2. Proofs.

LEMMA 1 (see [1]). *Let A be an $n \times n$ Hermitian matrix. Let $z \in \mathbb{C}^n$ be a normalized eigenvector of A corresponding to the eigenvalue λ. Denote by $\varphi(\mu) = \det(\mu I - A)$. Then*

$$\operatorname{adj}(\lambda I - A) = \varphi'(\lambda) z z^*.$$

When A is real symmetric, $z \in \mathbb{R}^n$.

LEMMA 2 (see [2, pp. 300, 316]). *Let T be the following $n \times n$ real symmetric tridiagonal matrix*

$$T = \begin{pmatrix} 0 & \beta_1 & & & & \\ \beta_1 & 0 & \beta_2 & & & \\ & \beta_2 & 0 & \ddots & & \\ & & \ddots & \ddots & \beta_{n-2} & \\ & & & \beta_{n-2} & 0 & \beta_{n-1} \\ & & & & \beta_{n-1} & 0 \end{pmatrix}, \tag{3}$$

where β's are nonzero real numbers. Let $\varphi_k(\lambda)$ be the characteristic polynomial of T_k, $k = 1, \dots, n$. Then

(a) $\varphi_k(\lambda) = \lambda \varphi_{k-1}(\lambda) - \beta_{k-1}^2 \varphi_{k-2}(\lambda)$, $k = 2, \dots, n$, *and* $\varphi_0(\lambda) \equiv 1$.

(b) *The eigenvalues of T are simple. The vector $x = (x_1, \dots, x_n)^T$ is an eigenvector of T corresponding to the eigenvalue λ, where*

$$x_1 = 1, \quad x_k = \frac{\varphi_{k-1}(\lambda)}{\beta_1 \dots \beta_{k-1}}, \qquad k = 2, \dots, n.$$

Proof of Theorem 1. (I) was already discussed in the previous section. We are going to establish (II) and (III).

($\Rightarrow$) Assume that the subsequences $\{x_{nr+k}\}$ are in arithmetic progression; i.e., $x_{nr+k} = x_{(n-1)r+k} + d_k$ for some $d_k \in \mathbb{R}$, where $k = 1, \dots, r$, $n = 1, 2, \dots$.

(II) When $r = 2$, we consider two cases.

Case 1. $d = (d_1, d_2) = 0$. The first three equations of (1) become

$$\begin{aligned} s_1 x_2 &= \lambda x_1, \\ s_1 x_1 + s_2 x_1 &= \lambda x_2, \\ s_2 x_2 + s_1 x_2 &= \lambda x_1, \end{aligned}$$

where $x_1 \neq 0$. We consider two possibilities.

(i) If $s_1 + s_2 \neq 0$, then from the second equation, $x_2 \neq 0$. Hence from the first and the third equations, we have $s_1 = s_1 + s_2$, which is impossible.

(ii) If $s_1 + s_2 = 0$, then from the third equation, $\lambda = 0$.

Case 2. $d \neq 0$. Subtracting the second from the fourth and the third from the fifth,

$$\begin{aligned} s_1 d_1 + s_2 d_1 &= \lambda d_2, \\ s_2 d_2 + s_1 d_2 &= \lambda d_1. \end{aligned}$$

Then $(s_1 + s_2)^2 d_2 = \lambda^2 d_2$ and $(s_1 + s_2)^2 d_1 = \lambda^2 d_1$. Since $d = (d_1, d_2) \neq 0$, we have $\lambda^2 = (s_1 + s_2)^2$. So $\lambda = \pm(s_1 + s_2)$.

Combining these two cases, $\lambda = \pm(s_1 + s_2)$ when $\{x_{2n}\}$ and $\{x_{2n+1}\}$ are in arithmetic progression.

(III) When $r \geq 3$, we consider two cases.

Case 1. $d = (d_1, \ldots, d_r)^T \neq 0$. Using the third set of r equations to subtract the second set of r equations in (1), we have

$$\begin{aligned} s_r d_r + s_1 d_2 &= \lambda d_1 \\ s_1 d_1 + s_2 d_3 &= \lambda d_2 \\ &\vdots \\ s_{r-1} d_{r-1} + s_r d_1 &= \lambda d_r. \end{aligned}$$

This means that λ is an eigenvalue of S' with eigenvector d.

Case 2. $d = 0$. Subtracting the $(r+1)$st equation of (1) from the first, we have $x_r = 0$. Then the first set of r equations in (1) becomes

$$\begin{aligned} s_1 x_2 &= \lambda x_1 \\ s_1 x_1 + s_2 x_3 &= \lambda x_2 \\ &\vdots \\ s_{r-1} x_{r-1} + s_r x_1 &= 0. \end{aligned}$$

Since $x_1 \neq 0$, λ is an eigenvalue of S' with eigenvector $x = (x_1, \ldots, x_{r-1}, 0)^T$.

($\Leftarrow$) (II) When $\lambda = \pm(s_1 + s_2)$, the relation (1) amounts to

$$\begin{aligned} s_1 x_2 &= \pm(s_1 + s_2) x_1, \\ s_1 x_{2n-1} + s_2 x_{2n+1} &= \pm(s_1 + s_2) x_{2n}, \quad n = 1, 2, \ldots, \\ s_2 x_{2n} + s_1 x_{2n+2} &= \pm(s_1 + s_2) x_{2n+1}, \quad n = 1, 2, \ldots . \end{aligned}$$

Combining the last two equations yields

$$\begin{aligned} x_{2n+2} - x_{2n} &= x_{2n+1} - x_{2n-1}, \qquad n = 1, 2, \ldots \quad \text{if} \quad \lambda = s_1 + s_2, \\ x_{2n+2} - x_{2n} &= x_{2n-1} - x_{2n+1}, \qquad n = 1, 2, \ldots \quad \text{if} \quad \lambda = -(s_1 + s_2), \end{aligned}$$

respectively. Similarly by shifting the index n to $n-1$ in the third equation and then combining with the second equation, we have

$$\begin{aligned} x_{2n+1} - x_{2n-1} &= x_{2n} - x_{2n-2}, \qquad n = 2, 3, \ldots \quad \text{if} \quad \lambda = s_1 + s_2, \\ x_{2n+1} - x_{2n-1} &= x_{2n-2} - x_{2n}, \qquad n = 2, 3, \ldots \quad \text{if} \quad \lambda = -(s_1 + s_2), \end{aligned}$$

respectively. So the two subsequences $\{x_{2n}\}$ and $\{x_{2n+1}\}$ are in arithmetic progression. In addition, $d_1 = \pm d_2$ if $\lambda = \pm(s_1 + s_2)$.

If $\lambda = s_1 + s_2$, then $x_3 - x_1 = x_2$ by considering the first two equations of (1). So $d_1 = d_2 = x_2 = \frac{s_1+s_2}{s_1} x_1$ and hence

$$x_{2n} = x_2 + (n-1)d_1 = x_2 + (n-1)(x_3 - x_1) = nx_2 = \frac{n(s_1+s_2)}{s_1} x_1, \quad n = 1, 2, \ldots,$$
$$x_{2n+1} = x_1 + nx_2 = \left(1 + \frac{n(s_1+s_2)}{s_1}\right) x_1, \quad n = 1, 2, \ldots.$$

If $\lambda = -(s_1 + s_2)$, then $x_2 = x_1 - x_3$ by considering the first two equations of (1). So $d_2 = -d_1 = x_2 = -\frac{s_1+s_2}{s_1} x_1$ and hence

$$x_{2n} = x_2 - (n-1)d_1 = x_2 - (n-1)(x_3 - x_1) = nx_2 = -\frac{n(s_1+s_2)}{s_1} x_1, \quad n = 1, 2, \ldots,$$
$$x_{2n+1} = x_1 - nx_2 = \left(1 + \frac{n(s_1+s_2)}{s_1}\right) x_1, \quad n = 1, 2, \ldots.$$

(III) Let λ be an eigenvalue of S'. We consider a modified relation:

$$\begin{aligned}
s_r x_0 + s_1 x_2 &= \lambda x_1 \\
s_1 x_1 + s_2 x_3 &= \lambda x_2 \\
&\vdots \\
s_{r-1} x_{r-1} + s_r x_{r+1} &= \lambda x_r \\
s_r x_r + s_1 x_{r+2} &= \lambda x_{r+1} \\
s_1 x_{r+1} + s_2 x_{r+3} &= \lambda x_{r+2} \\
&\vdots \\
s_{r-1} x_{2r-1} + s_r x_{2r+1} &= \lambda x_{2r} \\
&\vdots
\end{aligned} \tag{4}$$

The only difference between (1) and (4) is the introduction of the new variable x_0. When $x_0 = 0$, the system (1) is recovered. Set

$$d_j = x_{r+j} - x_j, \quad j = 0, 1, \ldots. \tag{5}$$

By subtracting the $(t+1)$st set of r equations of (4) from the $(t+2)$nd set of r equations, $t = 0, 1, \ldots$, the sequence $\{d_0, d_1, \ldots, d_r, d_{r+1}, d_{r+2}, \ldots\}$ satisfies (4).

Now we set $x_0 = 0$; i.e., $\{x_n\}$ is the sequence in (1) and $d_0 = x_r - x_0 = x_r$ and $d \equiv (d_1, \ldots, d_r)^T$. Then we make the following claim:[1]

$$S'd = \lambda d. \tag{6}$$

[1] In view of the proof of the implication ($\Rightarrow$) of Theorem 1, one can add the statement of the claim to (III) of Theorem 1. The proof will then be simpler because we can get around the proof of the claim. However, the present statement of (III) is weaker and hence serves better as a sufficient condition.

The claim (6) is equivalent to saying that d satisfies the following system of linear equations:

$$\begin{aligned} s_r z_r + s_1 z_2 &= \lambda z_1 \\ s_1 z_1 + s_2 z_3 &= \lambda z_2 \\ &\vdots \\ s_{r-1} z_{r-1} + s_r z_1 &= \lambda z_r. \end{aligned} \tag{7}$$

It follows that the periodic sequence $\hat{d} = \{d_r, d_1, d_2, \ldots, d_r, d_1, d_2, \ldots\}$ satisfies (4) as we duplicate (7) indefinitely. We recall that the sequence $\tilde{d} = \{d_0, d_1, \ldots, d_r, d_{r+1}, d_{r+2}, \ldots\}$ also satisfies (4), where d's are defined in (5), and $d_0 = x_r$. So the sequence $\{d_r - d_0, 0, 0, \ldots, d_1 - d_{r+1}, d_2 - d_{r+2}, \ldots, d_r - d_{2r}, \ldots\}$, which is the difference of $\hat{d}$ and $\tilde{d}$, satisfies (4). The zero terms of the sequence yield that

$$d_r = d_0 = x_r, \qquad d_{kr+j} = d_j, \quad j = 1, 2, \ldots, r, \quad k = 1, 2, \ldots \tag{8}$$

since $r \geq 3$. We now use induction on k to deduce that

$$x_{kr+j} = x_j + kd_j, \quad j = 1, 2, \ldots r, \quad k = 1, 2, \ldots.$$

The case $k = 1$ is obviously true because of (5) in which case we consider $j = 1, \ldots, r$. Now we assume that the statement is true for $k-1$, i.e., $x_{(k-1)r+j} = x_j + (k-1)d_j$, $j = 1, 2, \ldots, r$. Then using (5), the induction hypothesis, and (8), respectively, we get

$$\begin{aligned} x_{kr+j} &= x_{r+(k-1)r+j} \\ &= x_{(k-1)r+j} + d_{(k-1)r+j} \\ &= x_j + (k-1)d_j + d_j \\ &= x_j + kd_j, \qquad j = 1, 2, \ldots, r. \end{aligned}$$

Now we are going to prove the claim (6).

(a) Let λ be a simple eigenvalue of S' and let z be a normalized eigenvector of S' corresponding to λ.

In order to prove the claim (6), we first establish that (d_1, d_0) is a scalar multiple of (z_1, z_r). By Lemma 1, if $z = (z_1, \ldots, z_r)^T$ is a normalized eigenvector of S' corresponding to λ, then by considering the $(1, r)$ entry of $\operatorname{adj}(\lambda I - S')$, we have

$$\begin{aligned} z_1 z_r &= \frac{1}{\varphi'(\lambda)}(-1)^{r+1} \det \begin{pmatrix} -s_1 & & & & -s_r \\ \lambda & -s_2 & & & \\ -s_2 & \lambda & \ddots & & \\ & \ddots & \ddots & -s_{r-2} & \\ & & -s_{r-2} & \lambda & -s_{r-1} \end{pmatrix} \\ &= \frac{1}{\varphi'(\lambda)}(-1)^{r+1} \left\{(-1)^{r+1} s_r \varphi_{2,r-1}(\lambda) + (-1)^{r-1} s_1 \cdots s_{r-1}\right\} \qquad (9) \\ &= \frac{1}{\varphi'(\lambda)} \left\{s_r \varphi_{2,r-1}(\lambda) + s_1 \cdots s_{r-1}\right\} \end{aligned}$$

by Laplace expansion along the last column, where $\varphi_j(\lambda)$ and $\varphi_{i,j}(\lambda)$ are the characteristic polynomials of $S_j \equiv S_{1,j}$ and S_{ij}, respectively, and S is the matrix given in

(2) and $\varphi(\lambda) = \det(\lambda I - S')$. Notice that $\varphi'(\lambda) \neq 0$ as λ is a simple eigenvalue of S'. Similarly, by considering the (r, r) entry of $\operatorname{adj}(\lambda I - S')$, we have

$$z_r^2 = \frac{1}{\varphi'(\lambda)}\varphi_{r-1}(\lambda). \tag{10}$$

Without loss of generality, we may assume that $x_1 = 1$. By Lemma 2(b),

$$x_{r+1} = \frac{\varphi_r(\lambda)}{s_1 \cdots s_r}.$$

So

$$d_1 = x_{r+1} - x_1 = \frac{\varphi_r(\lambda)}{s_1 \cdots s_r} - 1 \tag{11}$$

and

$$d_0 = x_r = \frac{\varphi_{r-1}(\lambda)}{s_1 \cdots s_{r-1}}. \tag{12}$$

Since λ is an eigenvalue of S', by Laplace expansion along the last row, we have

$$\begin{aligned}
0 &= \varphi(\lambda) \\
&\equiv \det(\lambda I - S') \\
&= \det \begin{pmatrix}
\lambda & -s_1 & & & & -s_r \\
-s_1 & \lambda & -s_2 & & & \\
& -s_2 & \lambda & \ddots & & \\
& & \ddots & \ddots & -s_{r-2} & \\
& & & -s_{r-2} & \lambda & -s_{r-1} \\
-s_r & & & & -s_{r-1} & \lambda
\end{pmatrix} \\
&= (-1)^{r+1}(-s_r)\det \begin{pmatrix}
-s_1 & & & & -s_r \\
\lambda & -s_2 & & & \\
-s_2 & \lambda & \ddots & & \\
& \ddots & \ddots & -s_{r-2} & \\
& & -s_{r-2} & \lambda & -s_{r-1}
\end{pmatrix} \\
&\quad + (-1)^{2r-1}(-s_{r-1})\det \begin{pmatrix}
\lambda & -s_1 & & & & -s_r \\
-s_1 & \lambda & -s_2 & & & \\
& -s_2 & \lambda & \ddots & & \\
& & \ddots & \ddots & -s_{r-3} & \\
& & & & \lambda & \\
& & & & -s_{r-2} & -s_{r-1}
\end{pmatrix} + \lambda\,\varphi_{r-1}(\lambda).
\end{aligned}$$

Laplace expansion along the last column for the two determinants yields

$$\begin{aligned}
0 &= (-1)^{r+1}(-s_r)\{(-1)^{r+1}s_r\varphi_{2,r-1}(\lambda) + (-1)^{r-1}s_1 \cdots s_{r-1}\} \\
&\qquad + s_{r-1}\{(-1)^r(-s_r)(-1)^{r-2}s_1 \cdots s_{r-2} - s_{r-1}\varphi_{r-2}(\lambda)\} + \lambda\,\varphi_{r-1}(\lambda) \\
&= -s_r^2\varphi_{2,r-1}(\lambda) - s_1 \cdots s_r - s_1 \cdots s_r - s_{r-1}^2\varphi_{r-2}(\lambda) + \lambda\,\varphi_{r-1}(\lambda) \\
&= -s_r^2\varphi_{2,r-1}(\lambda) - s_{r-1}^2\varphi_{r-2}(\lambda) + \lambda\,\varphi_{r-1}(\lambda) - 2s_1 \cdots s_r.
\end{aligned}$$

Hence

$$\lambda\,\varphi_{r-1}(\lambda) - s_{r-1}^2\varphi_{r-2}(\lambda) = s_r^2\varphi_{2,r-1}(\lambda) + 2s_1\cdots s_r.$$

By Lemma 2(a), we have

$$\varphi_r(\lambda) = \lambda\varphi_{r-1}(\lambda) - s_{r-1}^2\varphi_{r-2}(\lambda) = s_r^2\varphi_{2,r-1}(\lambda) + 2s_1\cdots s_r.$$

Hence from (11)

$$d_1 = \frac{s_r\varphi_{2,r-1}(\lambda) + s_1\cdots s_{r-1}}{s_1\cdots s_{r-1}}.$$

Then from (9), (10), (11), and (12), we have

$$(d_1, d_0) = \frac{1}{s_1\cdots s_{r-1}}(s_r\varphi_{2,r-1}(\lambda) + s_1\cdots s_{r-1}, \varphi_{r-1}(\lambda)) = \frac{\varphi'(\lambda)z_r}{s_1\cdots s_{r-1}}(z_1, z_r). \tag{13}$$

Now we make the observation that the sequence defined by (4) is uniquely determined by x_0 and x_1. Indeed, the system (4) determines $x_2, x_3, \ldots$ as linear functionals[2] of the ordered pair (x_0, x_1).

(i) If $z_r = 0$, then by (13), $d_1 = d_0 = 0$. The sequence $\{d_0, d_1, \ldots, d_r, d_{r+1}, \ldots\}$, which satisfies (4), becomes the constant sequence $\{0, 0, \ldots\}$. Hence $\{x_{nr+k}\} = \{x_k\}$, $k = 1, \ldots, r-1$, and $x_{nr} = 0$, $n = 1, 2, \ldots$.

(ii) Suppose that $z_r \neq 0$. Now if z_0 is defined as z_r, then z_0 and z_1 will generate the periodic sequence $\{z_r, z_1, \ldots, z_{r-1}, z_r, z_1, \ldots\}$ via (4). Hence the ordered pair (d_1, d_0), a nonzero multiple of (z_1, z_r), generates the periodic sequence $\{d_0, d_1, \ldots, d_{r-1}, d_r, d_1, \ldots\}$, where $d_0 = d_r$. Moreover, $d = \varphi'(\lambda)z_r(s_1\cdots s_{r-1})^{-1}z$ since λ is a simple eigenvalue of S'.

If $\lambda \equiv \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_r$ are the eigenvalues of S',

$$\begin{aligned} d_r = d_0 &= \frac{\varphi'(\lambda)}{s_1\cdots s_{r-1}}z_r^2 \\ &= \frac{(\lambda_1-\lambda_2)\ldots(\lambda_1-\lambda_r)z_r^2}{s_1\cdots s_{r-1}} \\ &> 0. \end{aligned}$$

(b) In order to prove (b), it is sufficient to show that $x_r = 0$ and $x_{r+1} = 1$ if $x_1 = 1$. If λ is a double root of $\det(\lambda I - S') = 0$, then according to Lemma 1, $\operatorname{adj}(\lambda I - S') = 0$. By considering the (r, r) and the $(r-1, r)$ entries of $\operatorname{adj}(\lambda I - S')$, respectively, we have

$$\varphi_{r-1}(\lambda) = 0, \tag{14}$$

$$s_1\cdots s_r + s_{r-1}^2\varphi_{r-2}(\lambda) = 0. \tag{15}$$

Indeed the expressions (14) and (15) appeared in the computation between (11) and (13). By Lemma 2(a), (14), and (15), we have

$$\varphi_r(\lambda) = \lambda\varphi_{r-1}(\lambda) - s_{r-1}^2\varphi_{r-2}(\lambda) = s_1\cdots s_r.$$

[2]In particular, $x_{r+j} = f_j(x_0, x_1)$, $j = 1, 2, \ldots, r$, where f's are linear functionals. Due to the periodic recursive relation of (4), we have $x_{kr+j} = f_j(x_{(k-1)r}, x_{(k-1)r+1})$, $j = 1, 2, \ldots, r$, $k = 1, 2, \ldots$.

So by Lemma 2(b) and (14), we have

$$x_r = \frac{\varphi_{r-1}(\lambda)}{s_1 \cdots s_{r-1}} = 0.$$

By Lemma 2(b), we have

$$x_{r+1} = \frac{\varphi_r(\lambda)}{s_1 \cdots s_r} = 1.$$

If we denote by $\lambda_1 \geq \cdots \geq \lambda_r$ the eigenvalues of S' and by $\mu_1 > \cdots > \mu_{r-1}$ (Lemma 2(b)) the eigenvalues of S_{r-1}, the interlacing inequalities [2] for the real symmetric S' are stated as

$$\lambda_1 \geq \mu_1 \geq \lambda_2 \geq \cdots \geq \mu_{r-1} \geq \lambda_r.$$

So the multiplicities of λ's are either 1 or 2. □

Proof of Corollary 1. By Lemma 2(b)

$$x_{nr+k} = \frac{\varphi_{nr+k-1}(\lambda)}{(s_1 \cdots s_r)^n s_1 \cdots s_{k-1}}, \quad k = 1, \ldots, r-1, \quad n = 1, 2, \ldots,$$
$$x_{nr} = \frac{\varphi_{nr-1}(\lambda)}{(s_1 \cdots s_r)^n s_1 \cdots s_{r-1}}, \quad n = 1, 2, \ldots.$$

(a) By (a) (ii) of Theorem 1 (III), we have $x_{nr+k} = x_k + nd_k$, where $d_k = \frac{\varphi'(\lambda) z_r z_k}{s_1 \cdots s_{r-1}}$, $k = 1, \ldots, r-1$, and $x_{nr} = nx_r$ for all $n = 1, 2, \ldots$. So we have

$$\frac{\varphi_{nr+k-1}(\lambda)}{(s_1 \cdots s_r)^n s_1 \cdots s_{k-1}} = \frac{\varphi_{k-1}(\lambda)}{s_1 \cdots s_{k-1}} + n\frac{\varphi'(\lambda) z_r z_k}{s_1 \cdots s_{r-1}}, \quad k = 1, \ldots, r-1, \quad n = 1, 2, \ldots,$$
$$\frac{\varphi_{nr-1}(\lambda)}{(s_1 \cdots s_r)^n s_1 \cdots s_{r-1}} = \frac{n\varphi_{r-1}(\lambda)}{s_1 \cdots s_{r-1}}, \quad n = 1, 2, \ldots.$$

Hence we have

$$\varphi_{nr+k-1}(\lambda) = s_1 \cdots s_r \varphi_{k-1}(\lambda) + n\frac{(s_1 \cdots s_r)^n \varphi'(\lambda) z_r z_k}{s_k \cdots s_{r-1}},$$

where $\varphi_0(\lambda) \equiv 1$, $k = 1, \ldots, r-1$, $n = 1, 2, \ldots$, and

$$\varphi_{nr-1}(\lambda) = n(s_1 \cdots s_r)^{n-1} \varphi_{r-1}(\lambda), \qquad n = 1, 2, \ldots.$$

(b) By (a) (i) and (b) of Theorem 1 (III), we have $x_{nr+k} = x_k$, $k = 1, \ldots, r-1$, and $x_{nr} = 0$ for all $n = 1, 2, \ldots$. It follows that

$$\varphi_{nr+k-1}(\lambda) = (s_1 \cdots s_r)^n \varphi_{k-1}(\lambda), \qquad k = 1, \ldots, r-1, \quad n = 1, 2, \ldots,$$

where $\varphi_0(\lambda) \equiv 1$, and

$$\varphi_{nr-1}(\lambda) = 0, \qquad n = 1, 2, \ldots. \qquad \square$$

3. Generalizations and examples.

Remark 1. If all s's are positive, then S' is a nonnegative irreducible symmetric matrix. The largest eigenvalue, denoted by λ, is the Perron root of S' and the Perron vector is positive. So, corresponding to this λ, the d's are all positive because $d_r > 0$ by Theorem 1 (III)(a)(ii).

By using MATLAB, we compute some numerical examples.

Example 1. If $r = 3$ and $s_1 = 1$, $s_2 = 2$, and $s_3 = 2$, the eigenvalues of

$$S' = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 2 \\ 2 & 2 & 0 \end{pmatrix}$$

are $\lambda_1 = -1$, $\lambda_2 = -2.3723$, $\lambda_3 = 3.3723$. They are simple eigenvalues. The corresponding eigenvectors are $(1, -1, 0)^T$, $(0.4544, 0.4544, -0.7662)^T$, $(0.5418, 0.5418, 0.6426)^T$.

For $\lambda_1 = -1$, the sequence $\{x_n\}$ is

$$\{1, -1, 0, 1, -1, 0, 1, -1, 0, \ldots\}.$$

So the three subsequences are

$$\{1, 1, \ldots\}, \quad \{-1, -1, \ldots\}, \quad \{0, 0, \ldots\}.$$

Notice that $z_3 = 0$; i.e., $d_3 = 0$, and this demonstrates (III)(a)(i) of Theorem 1.

For $\lambda_2 = -2.3723$, the sequence $\{x_n\}$ is

$$\{1, -2.3723, 2.3139, -0.3723, -3.7446, 4.6277, -1.7446, -5.1168, 6.9416, -3.1168, \\ -6.4891, 9.2554, -4.4891, -7.8614, \ldots\}.$$

So the three subsequences are

$$\{1, 1 + d_1, \ldots\}, \quad \{-2.3723, -2.3723 + d_2, \ldots\}, \quad \{2.3139, 4.6278, \ldots\},$$

where $d_3 = x_3 = 2.3139$, $d_1 = d_2 = \frac{0.4544}{-0.7662}(2.3139) = -1.3723$.

For $\lambda_3 = 3.3723$ (the Perron root of S'), the sequence $\{x_n\}$ is

$$\{1, 3.3723, 5.1861, 5.3723, 7.7446, 10.3723, 9.7446, 12.1168, 15.5584, 14.1168, \\ 16.4891, 20.7446, 18.4891, 20.8614, \ldots\}.$$

So the three subsequences are

$$\{1, 1 + d_1, \ldots\}, \quad \{3.3723, 3.3723 + d_2, \ldots\}, \quad \{5.1861, 10.3722, \ldots\},$$

where $d_3 = x_3 = 5.1861$, $d_1 = d_2 = \frac{0.5418}{0.6426}(5.1861) = 4.3723$. It is no wonder that d's are all positive because the Perron vector is positive.

Example 2 (double root case). (i) Let $s_1 = s_2 = 1$ and $s_3 = -1$ and $r = 3$. Then the eigenvalues of S' are $\lambda_1 = \lambda_2 = 1$ and $\lambda_3 = -2$. The corresponding eigenvectors are $(0.5910, 0.7834, 0.1924)^T$, $(0.5634, -0.2301, -0.7935)^T$, and $(1, -1, 1)^T$, respectively.

When $\lambda_1 = \lambda_2 = 1$ (double root), the sequence $\{x_n\}$ is

$$\{1, 1, 0, 1, 1, 0, \ldots\}.$$

So the three subsequences are

$$\{1, 1, \ldots\}, \quad \{1, 1, \ldots\}, \quad \{0, 0, \ldots\},$$

i.e., $d = 0$.

When $\lambda_3 = -2$, the sequence $\{x_n\}$ is

$$\{1, -2, 3, 4, -5, 6, 7, -8, 9, 10, -11, \ldots\}.$$

So the three subsequences are

$$\{1, 4, 7, \ldots\}, \quad \{-2, -5, -8, \ldots\}, \quad \{3, 6, 9, \ldots\},$$

i.e., $d = (3, -3, 3)$.

Remark 2. By direct computation, it can be shown that when $r = 3$, we have the cubic equation

$$\det(\lambda I - S') = \lambda^3 - (s_1^2 + s_2^2 + s_3^2)\lambda - 2s_1 s_2 s_3 = 0.$$

If s_1^2, s_2^2, and s_3^2 are not all equal (it is true when s_1, s_2, and s_3 are positive), then the roots of the above equation are real and distinct (see [3, p. 9]) since

$$\left(\frac{s_1^2 + s_2^2 + s_3^2}{3}\right)^3 > s_1^2 s_2^2 s_3^2$$

by arithmetic–geometric mean inequality. So the double root case, i.e., (c) of Theorem 1, does not occur when $r = 3$ if s_1, s_2, and s_3 are positive. The next example is corresponding to $r = 4$ and some positive s's.

Example 3 (double root case). Let $s_1 = 1, s_2 = 2, s_3 = 2, s_4 = 1$, and $r = 4$. Then the eigenvalues of S' are $\lambda_1 = \lambda_2 = 0$, $\lambda_3 = -3.1623$, $\lambda_4 = 3.1623$. The corresponding eigenvectors are $(0.6325, 0.5000, -0.3162, -0.5000)^T$, $(-0.6325, 0.5000, 0.3162, -0.5000)^T$, $(0.3162, -0.5000, 0.6325, -0.5000)^T$, and $(0.3162, 0.5000, 0.6325, 0.5000)^T$, respectively.

When $\lambda_1 = \lambda_2 = 0$ (double root), the sequence $\{x_n\}$ is

$$\{1, 0, -0.5, 0, 1, 0, -0.5, 0, \ldots\};$$

i.e., $d = 0$.

When $\lambda_3 = -3.1623$, the sequence $\{x_n\}$ is

$$\{1, -3.1623, 4.5000, -3.9528, 3.5000, -7.1151, 9.5000, -7.9057, 6.0000, -11.0680, 14.5000, \\ -11.8585, 8.5000, -15.0208, 19.5000, -15.8114, 11.0000, -18.9737, 24.5000, \ldots\};$$

i.e., $d = (2.5, -3.9528, 5, -3.9528)$.

When $\lambda_4 = 3.1623$, the sequence $\{x_n\}$ is

$$\{1, 3.1623, 4.5000, 3.9528, 3.5000, 7.1151, 9.5000, 7.9057, 6.0000, 11.0680, 14.5000, \\ 11.8585, 8.5000, 15.0208, 19.5000, 15.8114, 11.0000, 18.9737, 24.5000, \ldots\};$$

i.e., $d = (2.5, 3.9528, 5, 3.9528)$.

Remark 3. For a given positive integer m, we can duplicate the nonzero real numbers $s_1, \ldots, s_r$ m times, i.e., $s_1, \ldots, s_r, s_1, \ldots, s_r, \ldots, s_1, \ldots, s_r$, in which there are m sets of $s_1, \ldots, s_r$. Theorem 1 holds for the mr subsequences $\{x_{nmr+k}\}$, $k =$

$1, 2, \ldots, mr$, of $\{x_n\}$ and S' will be replaced by the $mr \times mr$ matrix $S'(m) = S_{mr} + s_r e_1 e_{mr}^T + s_r e_{mr} e_1^T$, where $e_1, \ldots, e_{mr}$ form the standard basis of $\mathbb{R}^{mr}$.

COROLLARY 2. *Let $s_1, \ldots, s_r$ be r given nonzero real numbers. Let $\{x_n\}$ be the sequence defined by* (1) *with $x_1 \neq 0$. Let $S'(m)$ be the matrix*

$$S_{mr} + s_r e_1 e_{mr}^T + s_r e_{mr} e_1^T.$$

Let m be a given positive integer.

Case 1. If $mr \leq 2$, i.e., $(m, r) = (1, 1)$, $(1, 2)$, or $(2, 1)$, then the statements for the first two cases are in Theorem 1. If $(m, r) = (2, 1)$, then the sequences $\{x_{2n}\}$ and $\{x_{2n+1}\}$ are in arithmetic progression if and only if $\lambda = \pm 2s_1$. If $\lambda = 2s_1$, we have $x_n = nx_1$, $n = 1, 2, \ldots$. If $\lambda = -2s_1$, we have $x_n = (-1)^{n+1} n x_1$, $n = 1, 2, \ldots$.

Case 2. If $mr \geq 3$, i.e, $m \geq 3$ and $r = 1$, or $(m, r) \geq (2, 2)$, then the subsequences $\{x_{nmr+k}\}$, $k = 1, \ldots, r$, are in arithmetic progression; i.e., $x_{nmr+k} = x_{(n-1)mr+k} + d_k$, $k = 1, \ldots, mr$, $n = 1, 2, \ldots$ if and only if λ is an eigenvalue of $S'(m)$. Moreover,

(a) if λ is a simple eigenvalue of $S'(m)$, then the vector $d = (d_1, \ldots, d_{mr})^T$ is in the null space of $S'(m) - \lambda I$ with $d_{mr} = x_{mr}$. In particular $\{x_{nmr}\} = \{nx_{mr}\}$. More precisely, if z is a normalized eigenvector of $S'(m)$ corresponding to λ, and

(i) if $z_{mr} = 0$, then $d = 0$, i.e., $x_{nmr+k} = x_k$, $k = 1, \ldots, mr - 1$, and $x_{nmr} = 0$ for all positive integers n. In other words, $\{x_n\}$ is a periodic sequence.

(ii) if $z_{mr} \neq 0$, then

$$d = \frac{\varphi'(\lambda) s_r z_{mr}}{(s_1 \cdots s_r)^m} z,$$

where $\varphi(\lambda)$ is the characteristic polynomial of $S'(m)$. Hence d is an eigenvector of $S'(m)$ with $d_{mr} = x_{mr}$. Moreover, if λ is the largest eigenvalue, then $d_{mr} > 0$.

(b) If λ is an eigenvalue of $S'(m)$ of multiplicity 2, then $d = 0$, i.e., $x_{nmr+k} = x_k$, $k = 1, \ldots, mr - 1$, and $x_{nmr} = 0$ for all positive integers n. In other words, $\{x_n\}$ is a periodic sequence.

There are no eigenvalues of $S'(m)$ with multiplicity greater than 2.

Example 4. Let $(m, r) = (3, 1)$ and $s_1 = 1$. Then

$$S' = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

has eigenvalues $0, 0, -2, 2$. Notice that the eigenvalue 0 has multiplicity 2. So by Case 2(b) of Corollary 2, we have $\{1, 0, -1, 0, 1, 0, -1, 0, \ldots\}$. This was the example we discussed at the very beginning.

COROLLARY 3. *Let $r \geq 3$. The spectrum of $S'(1)$ is a subset of the spectrum of $S'(m)$ for $m = 1, 2, \ldots$.*

Proof. Let λ be an eigenvalue of S'. By Theorem 1, the subsequences $\{x_{nr+k}\}$, $k = 1, \ldots, r$, corresponding to λ are in arithemtic progression. So are the $\{x_{nmr+k}\}$. Hence, by Corollary 2, λ is an eigenvalue of $S'(m)$. □

Example 5. Let $s_1 = 1$, $s_2 = 2$, $s_3 = 3$, $r = 3$, and $m = 2$. Then

$$S'(2) = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 3 \\ 1 & 0 & 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 2 \\ 3 & 0 & 0 & 0 & 2 & 0 \end{pmatrix}$$

and the eigenvalues of $S'(2)$ are

$$\lambda_1 = 0.9112, \lambda_2 = -0.9112, \lambda_3 = 3.2019, \lambda_4 = 4.1131, \lambda_5 = -3.2019, \lambda_6 = -4.1131.$$

The corresponding eigenvectors are

$$\begin{aligned}
&(0.3711, 0.5958, 0.0859, -0.3711, -0.5958, -0.0859)^T,\\
&(0.3711, -0.5958, 0.0859, 0.3711, -0.5958, 0.0859)^T,\\
&(0.4319, -0.1933, -0.5254, -0.4319, 0.1933, 0.5254)^T,\\
&(-0.4192, -0.3282, -0.4653, -0.4192, -0.3282, -0.4653)^T,\\
&(-0.4319, -0.1933, 0.5254, -0.4319, -0.1933, 0.5254)^T,\\
&(0.4192, -0.3282, 0.4653, -0.4192, 0.3282, -0.4653)^T.
\end{aligned}$$

When $\lambda_1 = 0.9112$, the sequence $\{x_n\}$ is

$$\{1, 0.9112, -0.0849, -0.6332, -0.3224, 0.1698, 0.2665, -0.2665, -0.2546, 0.1003,\\ 0.8553, 0.3395, -0.4671, -1.4441, -0.4244, 0.8338, 2.0329, 0.5093, -1.2006, -2.6217, \ldots\}.$$

When $\lambda_2 = -0.9112$, the sequence $\{x_n\}$ is

$$\{1, -0.9112, -0.0849, 0.6332, -0.3224, -0.1698, 0.2665, 0.2665, -0.2546, -0.1003,\\ 0.8553, -0.3395, -0.4671, 1.4441, -0.4244, -0.8338, 2.0329, -0.5093, -1.2006, 2.6217, \ldots\}.$$

When $\lambda_3 = 3.2019$, the sequence $\{x_n\}$ is

$$\{1, 3.2019, 4.6261, 2.8029, -4.9038, -9.2522, -6.6057, 6.6057, 13.8784,\\ 10.4086, -8.3076, -18.5045, -14.2115, 10.0096, 23.1306, 18.0143, -11.7115, -27.7567,\\ -21.8172, 13.4134, \ldots\}.$$

When $\lambda_4 = 4.1131$, the sequence $\{x_n\}$ is

$$\{1, 4.1131, 7.9588, 8.1696, 9.7262, 15.9175, 15.3393, 15.3393, 23.8763,\\ 22.5089, 20.9524, 31.8350, 29.6785, 26.5655, 39.7938, 36.8482, 32.1785, 47.7525,\\ 44.0178, 37.7916, \ldots\}.$$

When $\lambda_5 = -3.2019$, the sequence $\{x_n\}$ is

$$\{1, -3.2019, 4.6261, -2.8029, -4.9038, 9.2522, -6.6057, -6.6057, 13.8784, -10.4086,\\ -8.3076, 18.5045, -14.2115, -10.0096, 23.1306, -18.0143, -11.7115, 27.7567,\\ -21.8172, -13.4134, \ldots\}.$$

When $\lambda_6 = -4.1131$, the sequence $\{x_n\}$ is

$$\{1, -4.1131, 7.9588, -8.1696, 9.7262, -15.9175, 15.3393, -15.3393, 23.8763, -22.5089,\\ 20.9524, -31.8350, 29.6785, -26.5655, 39.7938, -36.8482, 32.1785, -47.7525,\\ 44.0178, -37.7916, \ldots\}.$$

The vector $d = (d_1, \ldots, d_6)$ can be obtained by direct computation for each case. We notice that λ_2, λ_4, and λ_5 are the eigenvalues of $S' \equiv S'(1)$.

Remark 4. For $r \geq 3$, one can consider complex s's; i.e.,

$$S = \begin{pmatrix} 0 & s_1 & & & & & & \\ \overline{s}_1 & 0 & s_2 & & & & & \\ & \overline{s}_2 & \ddots & \ddots & & & & \\ & & \ddots & 0 & s_{r-1} & & & \\ & & & \overline{s}_{r-1} & 0 & s_r & & \\ & & & & \overline{s}_r & 0 & s_1 & \\ & & & & & \overline{s}_1 & 0 & \ddots \\ & & & & & & \ddots & \ddots \end{pmatrix} \tag{16}$$

and

$$S' = \begin{pmatrix} 0 & s_1 & & & \overline{s}_r \\ \overline{s}_1 & 0 & s_2 & & \\ & \overline{s}_2 & \ddots & \ddots & \\ & & \ddots & 0 & s_{r-1} \\ s_r & & & \overline{s}_{r-1} & 0 \end{pmatrix}. \tag{17}$$

Lemma 1 works for Hermitian matrices. Lemma 2(a) can be modified as

$$\varphi_k(\lambda) = \lambda\,\varphi_{k-1}(\lambda) - |\beta_{k-1}|^2\varphi_{k-2}(\lambda), \qquad k = 2, \ldots, n, \qquad \varphi_0(\lambda) \equiv 1,$$

and Lemma 2(b) and the interlacing inequalities are still valid.

If $s_1 \cdots s_r \in \mathbb{R}$, then $D^{-1}S'D = \hat{S}'$, where

$$\hat{S}' = \begin{pmatrix} 0 & |s_1| & & & & & \pm|s_r| \\ |s_1| & 0 & |s_2| & & & & \\ & |s_2| & \ddots & \ddots & & & \\ & & \ddots & & |s_{r-2}| & & \\ & & & |s_{r-2}| & 0 & |s_{r-1}| & \\ \pm|s_r| & & & & |s_{r-1}| & 0 & \end{pmatrix}, \tag{18}$$

depending on whether the product $s_1 \cdots s_r$ is positive or negative and D is the diagonal unitary matrix

$$D = \operatorname{diag}(e^{i\theta_1}, \ldots, e^{i\theta_r}),$$

where

$$\theta_i - \theta_{i+1} = \arg s_i, \quad i = 1, \ldots, r-1, \qquad \theta_r - \theta_1 = \pm\arg s_r. \tag{19}$$

If $U = \oplus_{i=1}^{\infty} D$, then $\hat{S}y = \lambda y$, where $y = U^{-1}x$, $\hat{S} = U^{-1}SU$, and

$$\hat{S} = \begin{pmatrix} 0 & |s_1| & & & & & & \\ |s_1| & 0 & |s_2| & & & & & \\ & |s_2| & \ddots & \ddots & & & & \\ & & \ddots & 0 & |s_{r-1}| & & & \\ & & & |s_{r-1}| & 0 & \pm|s_r| & & \\ & & & & \pm|s_r| & 0 & |s_1| & \\ & & & & & |s_1| & 0 & \ddots \\ & & & & & & \ddots & \ddots \end{pmatrix}. \tag{20}$$

We have the following result for the complex case.

THEOREM 2. *Let $s_1, \dots, s_r$ be r given nonzero complex numbers, not all real, with $r \geq 3$. Let $\{x_n\}$ be the sequence defined by $Sx = \lambda x$, where S is given in* (16) *and $x_1 \neq 0$. The subsequences $\{x_{nr+k}\}$, $k = 1, \dots, r$, are in arithemtic progression; i.e., $x_{nr+k} = x_{(n-1)r+k} + d_k$ if and only if the products $s_1 \cdots s_r \in \mathbb{R}$ and λ is an eigenvalue of $\hat{S}'$ in* (18).

In addition, let θ's be given in (19) *if $s_1 \cdots s_r \in \mathbb{R}$.*

(a) *If λ is a simple eigenvalue of $\hat{S}'$, then the vector $\hat{d} = (e^{-i\theta_1} d_1, \dots, e^{-i\theta_r} d_r)^T$ is in the null space of $\hat{S}' - \lambda I$ with $x_r = d_r$. In particular $\{x_{nr}\} = \{x_r\}$. More precisely, if z is a normalized eigenvector of $\hat{S}'$ corresponding to λ, and*

(i) *if $z_r = 0$, then $d = 0$, i.e., $x_{nr+k} = x_k$, $k = 1, \dots, r-1$, and $x_{nr} = 0$ for all positive integers n. In other words, $\{x_n\}$ is a periodic sequence.*

(ii) *If $z_r \neq 0$, then*

$$(e^{-i\theta_1} d_1, \dots, e^{-i\theta_r} d_r) = \frac{\varphi'(\lambda)\overline{z}_r}{s_1 \cdots s_{r-1}} z,$$

where $\varphi(\lambda)$ is the characteristic polynomial of $\hat{S}'$. Hence $\hat{d}$ is an eigenvector of $\hat{S}'$ with $x_r = d_r$. Moreover, if $\lambda\,(> 0)$ is the largest eigenvalue, then $e^{-i\theta_r} d_r > 0$.

(b) *If λ is an eigenvalue of $\hat{S}'$ of multiplicity 2, then $d = 0$, i.e., $x_{nr+k} = x_k$, $k = 1, \dots, r-1$, and $x_{nr} = 0$ for all positive integers n. In other words, $\{x_n\}$ is a periodic sequence.*

Proof. ($\Rightarrow$) If the subsequences $\{x_{nr+k}\}$, $k = 1, \dots, r$, are in arithemtic progression, i.e., $x_{nr+k} = x_{(n-1)r+k} + d_k$, $k = 1, \dots, r$, $n = 1, 2, \dots$, then λ is an eigenvalue of S' in (17), and the proof is similar to that of Theorem 1 (III). Moreover, $d = (d_1, \dots, d_r)^T$ is in the null space of $S' - \lambda I$; i.e., the claim (7) is true. We also have $d_r = d_0 = x_r$, which is part of (8).

Case 1. Suppose that λ is a simple eigenvalue of S'. Let $z = (z_1, \dots, z_r)^T$ be an eigenvector of S' associated with the eigenvalue λ. Then $d = \mu z$ for some nonzero scalar μ. Now perform the computation that is similar to that in between (9) and (13). It yields

$$d_1 = x_{r+1} - x_1 = \frac{\overline{s}_r \varphi_{2,r-1}(\lambda) + (\overline{s}_1 \cdots \overline{s}_r) s_r^{-1}}{s_1 \cdots s_{r-1}}, \quad d_r = d_0 = x_r = \frac{\varphi_{r-1}(\lambda)}{s_1 \cdots s_{r-1}}.$$

Moreover,

$$z_1 \overline{z}_r = \frac{1}{\varphi'(\lambda)} \{\overline{s}_r \varphi_{2,r-1}(\lambda) + s_1 \cdots s_{r-1}\}, \quad |z_r|^2 = \frac{1}{\varphi'(\lambda)} \varphi_{r-1}(\lambda).$$

(i) $z_r = 0$. Then $z_1 \overline{z}_r = 0$; i.e., $|s_r|^2 \varphi_{2,r-1}(\lambda) + s_1 \cdots s_r = 0$. But $\varphi_{2,r-1}(\lambda) = (\lambda - \mu_1) \cdots (\lambda - \mu_{r-2})$, where μ's are the eigenvalues of the Hermitian matrix $S'_{2,r-1}$ and λ is an eigenvalue of the Hermitian matrix S'. So $\varphi_{2,r-1}(\lambda)$ and hence the product $s_1 \cdots s_r$ is real.

(ii) $z_r \neq 0$. Then $d_r = \mu z_r$ implies that $\mu = \varphi'(\lambda)\overline{z}_r / s_1 \cdots s_r$. So $d_1 = \mu z_1$ implies that $\overline{s}_r \varphi_{2,r-1}(\lambda) + (\overline{s}_1 \cdots \overline{s}_r) s_r^{-1} = \overline{s}_r \varphi_{2,r-1}(\lambda) + s_1 \cdots s_{r-1}$. Hence $s_1 \cdots s_r$ is real.

Case 2. Suppose that λ is a double root. Then by considering the $(r, r-1)$ entry of $\operatorname{adj}(\lambda I - S')$, Lemma 1 yields

$$\overline{s}_1 \cdots \overline{s}_r + |s_{r-1}|^2 \varphi_{r-2}(\lambda) = 0.$$

This also implies that $s_1 \cdots s_r$ is real.

The implication ($\Leftarrow$) now follows from Theorem 1 and $\hat{S}y = \lambda y$, where $y = U^{-1}x$ and $\hat{S} = U^{-1}SU$. □

We remark that the statement of Corollary 3 is valid for S in (16) when $s_1 \cdots s_r$ is real.

We need to take care of some special cases; i.e., $r = 1, 2$.

THEOREM 3. *Let $\{x_n\}$ be the sequence defined by $Sx = \lambda x$, where S is given in (16). Assume that $s_1, \ldots, s_r$ are given nonzero complex numbers and are not all real.*

(1) *If $r = 1$, then the sequence $\{x_n\}$ is not in arithmetic progression.*

(2) *If $r = 2$, and*

(i) *if $\overline{s}_1 + s_2 \neq 0$, then at least one of the subsequences $\{x_{2n}\}$ and $\{x_{2n+1}\}$ is not in arithmetic progression.*

(ii) *If $\overline{s}_1 + s_2 = 0$, then the subsequences $\{x_{2n}\}$ and $\{x_{2n+1}\}$ are in arithmetic progression if and only if $\lambda = 0$. In addition, $x_{2n} = 0$ and $x_{2n+1} = x_1$ for all positive integers n.*

Proof. (1) When $r = 1$, we have

$$\begin{aligned} s_1 x_2 &= \lambda x_1 \\ \overline{s}_1 x_1 + s_1 x_3 &= \lambda x_2 \\ &\vdots \\ \overline{s}_1 x_{r-1} + s_1 x_{r+1} &= \lambda x_r \\ &\vdots \end{aligned} \tag{21}$$

where $s_1 \neq 0$. Suppose that $\{x_n\}$ is in arithmetic progression; i.e., $x_n = x_1 + (n-1)d$. Subtracting the second equation from the third, we have $(\overline{s}_1 + s_1)d = \lambda d$. If $d \neq 0$, then $\lambda = \overline{s}_1 + s_1$. Substituting into the second equation, we have $\overline{s}_1 x_1 + s_1(x_1 + 2d) = (\overline{s}_1 + s_1)(x_1 + d)$. This implies that $s_1 = \overline{s}_1$; i.e., s_1 is real.

If $d = 0$, then from the first equation, $\lambda = s_1$. But then, from the second equation, $\overline{s}_1 + s_1 = s_1$; i.e., $s_1 = 0$, which is impossible.

(2) When $r = 2$, we have

$$\begin{aligned} s_1 x_2 &= \lambda x_1 \\ \overline{s}_1 x_1 + s_2 x_3 &= \lambda x_2 \\ \overline{s}_2 x_2 + s_1 x_4 &= \lambda x_3 \\ \overline{s}_1 x_3 + s_2 x_5 &= \lambda x_4 \\ &\vdots \end{aligned}$$

Suppose that $\{x_{2n}\}$ and $\{x_{2n+1}\}$ are in arithemtic progression; i.e., $x_{2n+1} = x_1 + nd_1$ and $x_{2n} = x_2 + (n-1)d_2$. We consider two cases.

Case 1. $d = (d_1, d_2) = 0$. The first three equations of (1) become

$$\begin{aligned} s_1 x_2 &= \lambda x_1, \\ \overline{s}_1 x_1 + s_2 x_1 &= \lambda x_2, \\ \overline{s}_2 x_2 + s_1 x_2 &= \lambda x_1, \end{aligned}$$

where $x_1 \neq 0$. We consider the following two possibilities:

(a) If $\overline{s}_1 + s_2 \neq 0$, then from the second equation, $x_2 \neq 0$. Hence from the first and the third equations we have $s_1 = \overline{s}_2 + s_1$; i.e., $s_2 = 0$. This is impossible.

(b) If $\overline{s}_1 + s_2 = 0$, then from the third equation, $\lambda = 0$.

Case 2. $d \neq 0$. Subtracting the second from the fourth and the third from the fifth, we have

$$\overline{s}_1 d_1 + s_2 d_1 = \lambda d_2,$$
$$\overline{s}_2 d_2 + s_1 d_2 = \lambda d_1.$$

Then $|\overline{s}_1 + s_2|^2 d_2 = \lambda^2 d_2$ and $|\overline{s}_1 + s_2|^2 d_1 = \lambda^2 d_1$. Since $d = (d_1, d_2) \neq 0$, $\lambda^2 = |\overline{s}_1 + s_2|^2$. So $\lambda = \pm|\overline{s}_1 + s_2|$. Direct computation gives

$$x_2 = \frac{\lambda x_1}{s_1},$$
$$x_3 = \frac{(\lambda x_2 - \overline{s}_1 x_1)}{s_2} = \frac{(\lambda^2 - |s_1|^2)x_1}{s_1 s_2},$$
$$x_4 = \frac{\lambda x_3 - \overline{s}_2 x_2}{s_1} = \frac{\lambda(\lambda^2 - |s_1|^2 - |s_2|^2)x_1}{s_1^2 s_2},$$
$$x_5 = \frac{\lambda x_4 - \overline{s}_1 x_3}{s_2} = \frac{\{\lambda^2(\lambda^2 - 2|s_1|^2 - |s_2|^2) + |s_1|^4\}x_1}{s_1^2 s_2^2},$$
$$x_6 = \frac{\lambda x_5 - \overline{s}_2 x_4}{s_1} = \frac{\lambda\{\lambda^2(\lambda^2 - 2|s_1|^2 - 2|s_2|^2) + |s_1|^4 + |s_1|^2|s_2|^2 + |s_2|^4\}x_1}{s_1^3 s_2^2}.$$

Since $x_5 - x_3 = x_3 - x_1$, i.e., $x_5 + x_1 = 2x_3$, we have

$$\frac{\lambda^2(\lambda^2 - 2|s_1|^2 - |s_2|^2) + |s_1|^4}{s_1^2 s_2^2} + 1 = \frac{2(\lambda^2 - |s_1|^2)}{s_1 s_2};$$

i.e.,

$$\begin{aligned} 0 &= \lambda^2(\lambda^2 - 2|s_1|^2 - |s_2|^2) + |s_1|^4 + s_1^2 s_2^2 - 2s_1 s_2 \lambda^2 + 2s_1 s_2 |s_1|^2 \\ &= \lambda^4 - \lambda^2(2s_1 s_2 + 2|s_1|^2 + |s_2|^2) + |s_1|^4 + s_1^2 s_2^2 + 2s_1 s_2 |s_1|^2. \end{aligned} \tag{22}$$

Since $x_6 + x_2 = 2x_4$, we have

$$\frac{\lambda\{\lambda^2(\lambda^2 - 2|s_1|^2 - 2|s_2|^2) + |s_1|^4 + |s_1|^2|s_2|^2 + |s_2|^4\}}{s_1^3 s_2^2} + \frac{\lambda}{s_1} = 2\frac{\lambda(\lambda^2 - |s_1|^2 - |s_2|^2)}{s_1^2 s_2}.$$

If $\lambda \neq 0$, i.e., $\overline{s}_1 + s_2 \neq 0$, it amounts to

$$\begin{aligned} 0 &= \lambda^2(\lambda^2 - 2|s_1|^2 - 2|s_2|^2 - 2s_1 s_2) + |s_1|^4 + |s_1|^2|s_2|^2 + |s_2|^4 \\ &\quad + s_1^2 s_2^2 + 2s_1 s_2(|s_1|^2 + |s_2|^2) \\ &= \lambda^4 - \lambda^2(2|s_1|^2 + 2|s_2|^2 + 2s_1 s_2) + |s_1|^4 + |s_1|^2|s_2|^2 + |s_2|^4 \\ &\quad + s_1^2 s_2^2 + 2s_1 s_2(|s_1|^2 + |s_2|^2). \end{aligned} \tag{23}$$

Equating (22) and (23), we have $|s_2|^2\lambda^2 = |s_1|^2|s_2|^2 + |s_2|^4 + 2s_1 s_2 |s_2|^2$; i.e.,

$$\lambda^2 = |s_1|^2 + |s_2|^2 + 2s_1 s_2.$$

But $\lambda^2 = |\overline{s}_1 + s_2|^2 = |s_1|^2 + |s_2|^2 + \overline{s}_1 s_2 + s_1 \overline{s}_2$. So we have $2s_1 s_2 = \overline{s}_1 s_2 + s_1 \overline{s}_2$; i.e.,

$$s_1 s_2 = \operatorname{Re} \overline{s}_1 s_2. \tag{24}$$

So s_1s_2 is real, i.e., either (a) $s_1s_2 = |s_1||s_2|$ or (b) $s_1s_2 = -|s_1||s_2|$. Let $s_1 = |s_1|e^{i\theta}$.

(a) If $s_1s_2 = |s_1||s_2|$, then $s_2 = |s_2|e^{-i\theta}$. But (24) implies $|s_1||s_2| = \mathrm{Re}\,|s_1||s_2|e^{2i\theta}$. This amounts to $\theta = 0$ or π; i.e., s_1 and s_2 are real numbers and of the same sign.

(b) If $s_1s_2 = -|s_1|s_2|$, then $s_2 = |s_2|e^{i(\pi-\theta)}$. So $-|s_1||s_2| = \mathrm{Re}\,|s_1||s_2|e^{i(2\theta-\pi)}$. This has to be rejected. This amounts to $\theta = \pm\pi$; i.e., s_1 and s_2 are real numbers and of opposite signs. This is rejected, too.

Hence we proved (2)(i) and the implication ($\Rightarrow$) of (2)(ii) of Theorem 3. Moreover, if $\lambda = \overline{s}_1 + s_2 = 0$, then the system becomes

$$\begin{aligned} s_1x_2 &= 0, \\ \overline{s}_1x_{2n-1} + s_2x_{2n+1} &= 0, \quad n = 1, 2, \\ \overline{s}_2x_{2n} + s_1x_{2n+2} &= 0, \quad n = 1, 2, \ldots. \end{aligned}$$

So $x_{2n} = 0$ and $x_{2n+1} = x_1$, $n = 1, 2, \ldots$. This completes the proof of (2)(ii) of Theorem 3. □

Example 6. If $r = 3$ and $s_1 = 1 + i$, $s_2 = 1 - i$, and $s_3 = 2$, the eigenvalues of

$$S' = \begin{pmatrix} 0 & 1+i & 2 \\ 1-i & 0 & 1-i \\ 2 & 1+i & 0 \end{pmatrix}$$

are $\lambda_1 = 3.2361$, $\lambda_2 = -2$, and $\lambda_3 = -1.2361$. The corresponding eigenvectors are $(1, 0.6180 - 0.6180i, 1)^T$, $(1, 0, -1)^T$, and $(-0.3090 - 0.3090i, 1, -0.3090 - 0.3090i)^T$.

For $\lambda_1 = -1.2361$, the sequence $\{x_n\}$ is

$$\{1, -0.6180 - 0.6180i, -0.2361, 0.7639, -0.2361 - 0.2361i, -0.4721, 0.5279, 0.1459 + 0.1459i, -0.7082, 0.2918, 0.5279 + 0.5279i, -0.9443, 0.0557, 0.9098 + 0.9098i, \ldots\},$$

where $d_1 = -0.2361$, $d_2 = 0.3819 + 0.3819i$, and $d_3 = -0.2361$.

For $\lambda_2 = -2$, the sequence $\{x_n\}$ is

$$\{1, -1 - i, 1, 0, -1 - i, 2, -1, -1 - i, 3, -2, -1 - i, 4, -3, -1 - i, \ldots\},$$

where $d_1 = -1$, $d_2 = 0$, and $d_3 = 1$.

For $\lambda_3 = 3.2361$, the sequence $\{x_n\}$ is

$$\{1, 1.6180 + 1.6180i, 4.2361, 5.2361, 4.2361 + 4.2361i, 8.4721, 9.4721, 6.8541 + 6.8541i, 12.7082, 13.7082, 9.4721 + 9.4721i, 16.9443, 17.9443, 12.0902 + 12.0902i, \ldots\},$$

where $d_1 = 4.236$, $d_2 = 2.6181$, and $d_3 = 4.236$.

Acknowledgment. The author is thankful to the referee for the careful reading of the manuscript and for helpful suggestions.

REFERENCES

[1] R. C. Thompson and P. McEnteggert, *Principal submatrices* II, *The upper and lower quadratic inequalities*, Linear Algebra Appl., 1 (1968), pp. 211–243.
[2] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1988.
[3] W. H. Beyer, *Standard Mathematical Tables*, 27th ed., CRC Press, Boca Raton, FL, 1984.
[4] W. C. Ridge, *Numerical range of a weighted shift with periodic weights*, in Proc. Amer. Math. Soc., 55 (1976), pp. 107–110.
[5] T. Y. Tam, *On a conjecture of Ridge*, in Proc. Amer. Math. Soc., to appear.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 4, pp. 861–867, October 1997

005

A BOUND FOR THE MATRIX SQUARE ROOT WITH APPLICATION TO EIGENVECTOR PERTURBATION*

ROY MATHIAS†

Abstract. Let H be positive definite, let ΔH be Hermitian, and suppose that $\|H^{-1/2}(\Delta H)H^{-1/2}\| = 1$. It is shown that the least constant c_n such that

$$||(H+\eta\Delta H)^{1/2}H^{-1/2} - I|| \le c_n\eta + 0(\eta^2)$$

for all $n \times n$ H and ΔH grows like $\log n$. This fact has consequences in eigenvector perturbation theory.

Key words. matrix square root, eigenvector perturbation, Hadamard product

AMS subject classifications. 15A45, 47A60

PII. S50895479895295775

It is well known that taking square roots of scalars halves relative perturbations. In particular, if we fix t and ϵ with $|\epsilon/t| = 1$ and consider $\tilde{t} = t + \eta\epsilon$, then the relative perturbation in $\tilde{t}$ is η, while the relative perturbation in $\tilde{t}^{1/2}$, to first order, is

$$\left|\frac{\tilde{t}^{1/2} - t^{1/2}}{t^{1/2}}\right| = \left|\frac{\frac{1}{2}\eta\epsilon t^{-1/2}}{t^{1/2}}\right| = \frac{1}{2}\eta.$$

In this paper we consider two generalizations of this result to positive definite matrices. Let H and $\tilde{H} = H + \Delta H$ be $n \times n$ positive definite matrices. Since we must allow for noncommutativity it is natural to take the size of the relative perturbation in $\tilde{H}$ to be

$$\|H^{-1/2}(\Delta H)H^{-1/2}\|,$$

and to take the size of the relative perturbation in $\tilde{H}^{1/2}$ to be

$$\|H^{-1/4}(\tilde{H}^{1/2} - H^{1/2})H^{-1/4}\|.$$

Here, and unless otherwise stated, we use the spectral norm. By analogy with the scalar case one would hope that the following theorem would hold.

THEOREM 1. *Let H be positive definite and let $\tilde{H} = H + \eta\Delta H$, where ΔH is Hermitian and such that $\|H^{-1/2}(\Delta H)H^{-1/2}\| = 1$. Then*

$$\|H^{-1/4}(\tilde{H}^{1/2} - H^{1/2})H^{-1/4}\| \le \frac{1}{2}\eta + O(\eta^2). \tag{1}$$

Indeed it does, as we shall soon prove.

* Received by the editors December 4, 1995; accepted for publication (in revised form) by R. Bhatia September 20, 1996. This research was supported by grants from the National Science Foundation and a Summer Research Grant from the College of William and Mary.
http://www.siam.org/journals/simax/18-4/29577.html
† Department of Mathematics, College of William and Mary, Williamsburg, VA 23187 (mathias@math.wm.edu).

There are situations[1] where one wants to bound

$$\|\tilde{H}^{1/2}H^{-1/2} - I\| = \|(\tilde{H}^{1/2} - H^{1/2})H^{-1/2}\|, \tag{2}$$

in terms of $\|H^{-1/2}(\Delta H)H^{-1/2}\|$. Since for any matrices A and B, with B Hermitian, we have

$$\|ABA^*\| = \rho(ABA^*) = \rho(BA^*A) \le \|BA^*A\|$$

(here ρ denotes the spectral radius), it follows that

$$\|H^{-1/4}(\tilde{H}^{1/2} - H^{1/2})H^{-1/4}\| \le \|(\tilde{H}^{1/2} - H^{1/2})H^{-1/2}\| = \|\tilde{H}^{1/2}H^{-1/2} - I\|,$$

and so we should not be surprised if we cannot prove

$$\|\tilde{H}^{1/2}H^{-1/2} - I\| \le \frac{1}{2}\eta + O(\eta^2).$$

In fact, the best one can do is summarized in Theorem 2. The following constants occur in Theorem 2:

$$\gamma_n \equiv \frac{1}{n}\sum_{j=1}^{n} |\cot(2j-1)\pi/2n| = \frac{2}{n}\sum_{j=1}^{[n/2]} \cot(2j-1)\pi/2n$$

$$\hat{\gamma}_n \equiv \frac{1}{n}\sum_{j=1}^{n} |\csc(2j-1)\pi/2n|.$$

Since $|\cot\theta| < |\csc\theta| < |\cot\theta| + 1$ for $\theta \ne k\pi/2$, it follows that $\gamma_n < \hat{\gamma}_n < \gamma_n + 1$. Approximating the sums by integrals one can show that $\hat{\gamma}_n - \gamma_n \to (1/\pi)\log 2$, $\gamma_n/\log n \to 2/\pi$ and $\hat{\gamma}_n/\log n \to 2/\pi$, as $n \to \infty$. The sequences γ_n and $\hat{\gamma}_n$ are strictly increasing.

THEOREM 2. *Consider the bound*

$$\|(H + \eta\Delta H)^{1/2}H^{-1/2} - I\| \le c_n\eta + 0(\eta^2). \tag{3}$$

If this bound is to hold for all $n \times n$ positive definite matrices H and all Hermitian matrices ΔH such that $||H^{-1/2}(\Delta H)H^{-1/2}|| = 1$, then c_n must be at least $\frac{1}{2}(\gamma_n - 1)$.

On the other hand, (2) is valid for all $n \times n$ positive definite matrices H and all Hermitian ΔH with $\|H^{-1/2}(\Delta H)H^{-1/2}\| = 1$ provided that $c_n \ge \frac{3}{2} + \hat{\gamma}_{n-1}$.

To prove this result we first introduce some notation and then prove a preliminary lemma on Hadamard products that may be of independent interest.

Given matrices $A = [a_{ij}]_{i,j=1}^n$ and $B = [b_{ij}]_{i,j=1}^n$, we denote their *Hadamard product* (or componentwise product) by $A \circ B = [a_{ij}b_{ij}]_{i,j=1}^n$. The Schur product theorem states that if A and B are positive semidefinite, then so is $A \circ B$. Let $\lambda_i, \mu_i, i = 1, 2, \ldots, n$ be positive scalars; then $[(\lambda_i + \lambda_j)^{-1}]_{i,j=1}^n$ is a positive semidefinite

[1] For example, in relative perturbation theory one wants to bound the perturbation in the eigenvectors of $\tilde{H} = H + \Delta H$ in terms of $\|H^{-1/2}(\Delta H)H^{-1/2}\|$. One can obtain such a bound by writing $\tilde{H} = DHD^*$, where $D = \tilde{H}^{1/2}H^{-1/2}$, and then applying [3, Theorem 2.2]. The resulting bound includes the term $\|D - I\| = \|\tilde{H}^{1/2}H^{-1/2} - I\|$. Theorem 2 in this paper shows that if one uses [3, Theorem 2.2] in this way, then the resulting bound inevitably involves a factor of $\log n$. There are at least two ways to obtain a stronger eigenvector bound that does not contain a factor of $\log n$. One can more carefully use the method proof of [3, Theorem 2.1] and bound the quantity $\|(D - I)w_i\|$ that arises in the proof [2]. Alternatively one can use the main result in [7].

Cauchy matrix [4, Problem 5.5.9 (a), pp. 347–348], while $[\mu_i\mu_j]_{i,j=1}^n$ is a positive semidefinite rank-1 matrix. The Schur product theorem ensures that

$$\left[\frac{\mu_i\mu_j}{\lambda_i+\lambda_j}\right]_{i,j=1}^n$$

is positive semidefinite. Given a matrix A we define its *Hadamard operator norm*, denoted by $|||A|||$, by

$$|||A||| = \max\{||A\circ B|| : ||B|| \le 1\}.$$

For positive semidefinite A it is well known that (see, e.g., [4, Theorem 5.5.18, second part])

$$|||A||| = \max\{a_{11},\dots,a_{nn}\}.$$

Proof of Theorem 1. Without loss of generality we may assume that

$$H = \Lambda = \operatorname{diag}(\lambda_1,\dots,\lambda_n).$$

Then by assumption

$$A = \Lambda^{-1/2}\Delta H\Lambda^{-1/2}$$

has norm 1.

Now, using the so-called Daleckĭi–Kreĭn formula for the derivative of $(H+\eta\Delta H)^{1/2}$ with respect to η at $\eta = 0$ [4, Theorem 6.6.30 (1), noting that $U(0) = I$], we have

$$\begin{aligned}(H+\eta\Delta H)^{1/2} &= (\Lambda+\eta\Delta H)^{1/2}\\ &= \Lambda^{1/2} + \eta\left[\frac{\lambda_i^{1/2}-\lambda_j^{1/2}}{\lambda_i-\lambda_j}\right]_{i,j=1}^n \circ \Delta H + O(\eta^2)\\ &= \Lambda^{1/2} + \eta\left[\frac{1}{\lambda_i^{1/2}+\lambda_j^{1/2}}\right]_{i,j=1}^n \circ (\Lambda^{1/2}A\Lambda^{1/2}) + O(\eta^2)\\ &= \Lambda^{1/2} + \eta\left[\frac{\lambda_i^{1/2}\lambda_j^{1/2}}{\lambda_i^{1/2}+\lambda_j^{1/2}}\right]_{i,j=1}^n \circ A + O(\eta^2). \qquad (4)\end{aligned}$$

Thus to first order in η

$$\begin{aligned}\left\|H^{-1/4}\left[(H+\eta\Delta H)^{1/2}-H^{1/2}\right]H^{-1/4}\right\| &= \left\|\eta\Lambda^{-1/4}\left(\left[\frac{\lambda_i^{1/2}\lambda_j^{1/2}}{\lambda_i^{1/2}+\lambda_j^{1/2}}\right]_{i,j=1}^n \circ A\right)\Lambda^{-1/4}\right\|\\ &= \left\|\left[\frac{\lambda_i^{1/4}\lambda_j^{1/4}}{\lambda_i^{1/2}+\lambda_j^{1/2}}\right]_{i,j=1}^n \circ A\right\|\eta\\ &\le \frac{1}{2}\eta.\end{aligned}$$

For the final inequality we have used the fact that the matrix

$$\left[\frac{\lambda_i^{1/4}\lambda_j^{1/4}}{\lambda_i^{1/2}+\lambda_j^{1/2}}\right]_{i,j=1}^n$$

is positive semidefinite, and hence its Hadamard operator norm is its largest main diagonal element—that is, $\frac{1}{2}$. $\square$

Now let us turn our attention to Theorem 2. Assume from now on that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$. In proving Theorem 2 it is necessary to obtain bounds on

$$\left|\!\left|\!\left| \left[\frac{\lambda_i}{\lambda_i + \lambda_j} \right]_{i,j=1}^n \right|\!\right|\!\right|.$$

A simple approach to bounding this quantity is

$$\begin{aligned} \left|\!\left|\!\left| \left[\frac{\lambda_i}{\lambda_i + \lambda_j} \right]_{i,j=1}^n \right|\!\right|\!\right| &= \left|\!\left|\!\left| \left[\frac{\lambda_i \lambda_j}{\lambda_i + \lambda_j} \right]_{i,j=1}^n \circ \left[\frac{1}{\lambda_j} \right]_{i,j=1}^n \right|\!\right|\!\right| \\ &\leq \left|\!\left|\!\left| \left[\frac{\lambda_i \lambda_j}{\lambda_i + \lambda_j} \right]_{i,j=1}^n \right|\!\right|\!\right| \left|\!\left|\!\left| \left[\frac{1}{\lambda_j} \right]_{i,j=1}^n \right|\!\right|\!\right| \\ &= \frac{1}{2} \lambda_1 \frac{1}{\lambda_n}. \end{aligned} \tag{5}$$

For the final equality we have used the fact that $[\frac{\lambda_i \lambda_j}{\lambda_i + \lambda_j}]$ is positive semidefinite and hence that its Hadamard operator norm is its largest diagonal element [4, Theorem 5.5.18, second part]. The fact that $||| [\frac{1}{\lambda_j}] ||| = \frac{1}{\lambda_n}$ follows from

$$\left\| \left[\frac{1}{\lambda_j} \right]_{i,j=1}^n \circ B \right\| = ||B\Lambda^{-1}|| \leq ||\Lambda^{-1}||\, ||B|| = \lambda_n^{-1} ||B||,$$

where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$.

Typically we will apply the bound on $||| [\frac{\lambda_i}{\lambda_i + \lambda_j}] |||$ when H is graded and in this case $\frac{\lambda_1}{\lambda_n}$ is very large and so the bound (5) is not very useful, even though it is independent of n. Our next result gives bounds on $||| [\frac{\lambda_i}{\lambda_i + \lambda_j}] |||$ that are independent of $\frac{\lambda_1}{\lambda_n}$ but grow like $\log n$.

LEMMA 3. *Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$. Then*

$$\left|\!\left|\!\left| \left[\frac{\lambda_i}{\lambda_i + \lambda_j} \right] \right|\!\right|\!\right| \leq \frac{3}{2} + \hat{\gamma}_{n-1}. \tag{6}$$

Furthermore, one can choose $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ and a Hermitian contraction B such that $||| [\frac{\lambda_i}{\lambda_i + \lambda_j}] \circ B ||$ is at least $\frac{1}{2}(\gamma_n - 1) - \epsilon$ for any $\epsilon > 0$.

Proof. Let T_n denote the $n \times n$ strictly upper triangular matrix with all elements above the diagonal equal to 1. Then one can check that

$$\begin{aligned} \left[\frac{\lambda_i}{\lambda_i + \lambda_j} \right] &= \left[\frac{\min\{\lambda_i, \lambda_j\}}{\lambda_i + \lambda_j} \right] + T_n \circ \left[\frac{|\lambda_i - \lambda_j|}{\lambda_i + \lambda_j} \right] \\ &\equiv X + T_n \circ Y. \end{aligned}$$

We have used the fact that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ in the first equality. The matrix X is positive semidefinite—see, for example, the proof of Theorem 3.2 in [5], and so $|||X|||$ is just its largest diagonal element, which is $\frac{1}{2}$. The Hadamard operator norms of Y and T_n have been shown to be at most 2 [5, Theorem 3.2] and $(\hat{\gamma}_{n-1} + 1)/2$

([5, Corollary 3.5] and the fact that, in the notation of [5], $|||T_n||| = |||\hat{T}_{n-1}|||$), respectively. So

$$|||T_n \circ Y||| \le |||T_n||| \; |||Y||| \le 2 \cdot \frac{1}{2}(\hat{\gamma}_{n-1} + 1).$$

Combining these bounds gives (6).

Now let us consider the lower bound on $||||[\frac{\lambda_i}{\lambda_i+\lambda_j}]||||$. If we choose the λ's so that the ratio $\lambda_{i+1}/\lambda_i \to 0$ for each $i = 1, \ldots, n-1$, then

$$\left[\frac{\lambda_i}{\lambda_i + \lambda_j}\right] \to \tilde{T}_n \equiv T_n + \frac{1}{2}I.$$

Because $|||\cdot|||$ is continuous we need only show that there is a Hermitian contraction B such that

$$|||\tilde{T}_n \circ B||| \ge \frac{1}{2}(\gamma_n - 1).$$

Let $S_n = [\text{sign } (j-i)]$; that is, S_n has 0's on the diagonal, 1's above the diagonal, and -1's below the diagonal. Lemma 3.1 in [5] gives $|||S_n||| = \gamma_n$. There is a contraction B such that $||iS_n \circ B|| = |||iS_n|||$, and by [6, Corollary 3.3] we may take B to be a Hermitian contraction since iS_n is itself Hermitian. Then

$$||S_n \circ B|| = |||S_n||| = \gamma_n.$$

Let J_n denote the $n \times n$ matrix of 1's. One can check that $\tilde{T}_n = \frac{1}{2}(S_n + J_n)$. For the B in the previous paragraph we have

$$\begin{aligned} ||\tilde{T}_n \circ B|| &= \frac{1}{2}||S_n \circ B + J_n \circ B|| \\ &\ge \frac{1}{2}(||S_n \circ B|| - ||J_n \circ B||) \\ &= \frac{1}{2}(\gamma_n - ||B||) \\ &\ge \frac{1}{2}(\gamma_n - 1) \end{aligned}$$

as required. □

Now we can prove Theorem 2.

Proof of Theorem 2. As in the proof of Theorem 1 we may assume without loss of generality that

$$H = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n).$$

Then $A = \Lambda^{-1/2}\Delta H \Lambda^{-1/2}$ has norm-1 by assumption.

Now, using the first-order approximation in (4) we have

$$(H + \eta\Delta H)^{1/2}H^{-1/2} - I = \eta\left[\frac{\lambda_i^{1/2}}{\lambda_i^{1/2} + \lambda_j^{1/2}}\right] \circ A + 0(\eta^2).$$

Since the only constraint on ΔH is that $\|H^{-1/2}(\Delta H)H^{-1/2}\| = 1$, A is an arbitrary Hermitian matrix with norm 1. So the smallest value of c_n for which (3) is valid is

$$\alpha = \max\left\{\left\|\left[\frac{\lambda_i^{1/2}}{\lambda_i^{1/2} + \lambda_j^{1/2}}\right] \circ X\right\| : ||X|| \le 1, X = X^*\right\}.$$

From Lemma 3 we have

$$\alpha \le \left|\left|\left| \left[\frac{\lambda_i^{1/2}}{\lambda_i^{1/2}+\lambda_j^{1/2}} \right] \right|\right| \le \frac{3}{2} + \hat{\gamma}_{n-1},$$

which is the second part of the theorem. Lemma 3 also says that the λ's can be chosen so that α is arbitrarily close to $\frac{1}{2}(\gamma_n - 1)$, and this gives us the first part of the theorem. □

The bounds in this paper depend heavily on results on $|||[\text{sign}\,(j-i)]_{i,j=1}^n|||$ from [5]. See [5] for further applications of the fact $|||[\text{sign}\,(j-i)]_{i,j=1}^n||| = \gamma_n$ to inequalities for commutators and the matrix absolute value ($|A| \equiv (A^*A)^{1/2}$), to the norm of the triangular truncation operator, and to bounds on the perturbation of the eigenvalues of a Hermitian matrix that is subjected to a skew-Hermitian perturbation.

So far we have considered only the spectral norm. If we had taken the Frobenius norm,

$$||X||_F \equiv \left(\sum_{i,j} |x_{ij}|^2 \right)^{1/2},$$

then, since $|\lambda_i^{1/2}(\lambda_i^{1/2}+\lambda_j^{1/2})^{-1}| \le 1$, it is easy to see that

$$\left\| \left[\frac{\lambda_i^{1/2}}{\lambda_i^{1/2}+\lambda_j^{1/2}} \right] \circ A \right\|_F \le ||A||_F,$$

and hence that if $||\cdot||$ is replaced by $||\cdot||_F$ in (3) then we may take $c_n = 1$ independent of n.

Let $||\cdot||_p$ denote the Schatten p-norm, i.e.,

$$||X||_p = \left(\sum_{i=1}^n \sigma_i^p(x) \right)^{1/p},$$

and let

$$|||A|||_p \equiv \max\{||A \circ B||_p : ||B||_p \le 1\}.$$

Note that $||\cdot|| = ||\cdot||_\infty$ and $||\cdot||_F = ||\cdot||_2$. Davies [1, Proposition 4] has proved the following bounds that are independent of n for the Hadamard operator norm with respect to the Schatten p-norms:

$$\left|\left|\left| \left[\frac{\lambda_i - \lambda_j}{\lambda_i + \lambda_j} \right]_{i,j=1}^n \right|\right|\right| \le \begin{cases} cp, & 2 \le p < \infty, \\ cp/(1-p), & 1 < p \le 2, \end{cases}$$

where c is an absolute constant (independent of p, n, and λ). If we let $\lambda_{i+1}/\lambda_i \to 0$ as we did in the proof of Lemma 3, then

$$\left[\frac{\lambda_i - \lambda_j}{\lambda_i + \lambda_j} \right]_{i,j=1}^n \to \tilde{T}_n.$$

Thus $\left|\left|\left|\tilde{T}_n\right|\right|\right|$ is bounded independent of n and, following the proof of Lemma 3,

$$\left|\left|\left|\left[\frac{\lambda_i}{\lambda_i+\lambda_j}\right]\right|\right|\right|_p$$

is also bounded independently of n. Consequently, if we replace $||\cdot||$ by $||\cdot||_p$ in (1) for any $p \in (1,\infty)$, then we may choose c_n independent of n.

REFERENCES

[1] E. B. Davies, *Lipschitz continuity of functions of operators in the Schatten classes*, J. London Math. Soc., 37 (1988), pp. 148–157.

[2] S. Eisenstat, *personal communication*, 1996.

[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] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, London, 1991.

[5] R. Mathias, *The Hadamard operator norm of a circulant and applications*, SIAM J. Matrix Anal. Appl., 14 (1992), pp. 1152–1167.

[6] R. Mathias, *Matrix completions, Hadamard products and norms*, Proc. Amer. Math. Soc., 117 (1993), pp. 905–918.

[7] R. Mathias and K. Veselić, *A relative perturbation bound for positive definite matrices*, Linear Algebra Appl., to appear.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 4, pp. 868–886, October 1997

006

SMALL-SAMPLE STATISTICAL ESTIMATES FOR THE SENSITIVITY OF EIGENVALUE PROBLEMS*

THORKELL GUDMUNDSSON†‡, CHARLES KENNEY†, AND ALAN J. LAUB†

Abstract. In this paper a new approach to the evaluation of sensitivity or condition of eigenvalue problems is proposed. This approach is applicable to general nonsymmetric matrices as well as to matrices with special structure and is suitable for various types of perturbations. In particular, the important class of componentwise relative perturbations can easily be handled for a general matrix. This cannot be done satisfactorily with other currently available methods. The sensitivity evaluation is based on the recently introduced technique of small-sample statistical estimation for the local sensitivity of a large variety of functions.

Key words. conditioning, statistical condition estimation, eigenvalues, eigenvectors, sensitivity

AMS subject classifications. 65F35, 65F30, 65F15, 15A12

PII. S0895479895296021

1. Introduction.

The accuracy of numerical computations is traditionally measured in terms of an upper bound on the error in the computed result under the assumption that each data element is perturbed equally and that the overall perturbation is bounded in norm. Some examples are the well-known condition number of a matrix, which gives an upper bound on the norm of the deviation of a computed solution to a linear system from the correct solution [7], [22], [24], bounds on the error norm in a computed least-squares solution [41], and bounds on the error in computed results of eigenvalue problems. Sensitivity analysis for the latter is significantly more complex than for the first two, and many different approaches to measuring the condition of eigenproblems have been proposed. A classical treatment of the sensitivity of individual eigenvalues and eigenvector components of a general matrix under norm-bounded perturbations is [47]. In [8], the sensitivity of invariant subspaces of Hermitian matrices is defined in terms of rotations of the subspaces, and [40] extends this to general matrices. Further work on the behavior of ill-conditioned eigenvalues under perturbations may be found in [20] and [48], the sensitivity of eigenvector components is further discussed in [17], [35], and [36] focuses on Hermitian matrices. Extensions of these results have been made to the generalized eigenvalue problem [9], [39], [40], as well as to singular value problems [42], [46]. An excellent treatment of the eigenvalue problem and its sensitivity to perturbations is given in [5], while [19], [27], and [44] discuss the effects of norm-bounded perturbations on the both this and other matrix problems.

The assumption that each element of a matrix is perturbed equally in absolute terms is not appropriate in many practical situations and has led to increased research on other types of perturbations. For example, errors introduced by representing data in finite precision arithmetic might be modeled by assuming that each element of the matrix undergoes a random perturbation relative to its individual magnitude, rather

* Received by the editors December 8, 1995; accepted for publication (in revised form) by N. J. Higham September 23, 1996. This research was supported by Air Force Office of Scientific Research grant F49620-94-0104DEF, National Science Foundation grant ECS-9120643, and Icelandic Research Council grant 95-N-007.

http://www.siam.org/journals/simax/18-4/29602.html

† Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106-9560 (laub@ucdavis.edu).

‡ Engineering Research Institute, University of Iceland, Hjardarhagi 2-6, 107 Reykjavik, Iceland.

than relative to the norm of the matrix. The assumption of such componentwise relative perturbations in the matrix leads to very different upper bounds on the computational error, which are frequently much smaller than their counterparts based on norm-bounded errors. Although this difference may often be reduced by appropriate scaling of the matrix, the issue of choosing such scaling is not very well resolved. An excellent survey of current results on error bounds based on componentwise relative perturbations may be found in [25], and an extended treatment of this and many related topics can be consulted in [26].

Another important perturbation class is that of structured perturbations. This may arise when the matrix has some fixed structural integer entries such as coefficients of a discretized differential equation, fixed zeros such as in sparse or "shaped" matrices, or elements known to be equal or derived from a single parameter such as for symmetric or Vandermonde matrices. In the first two cases some elements are not perturbed at all, and in the third case not all elements are perturbed independently, so an analysis using norm-bounded perturbations is not appropriate. Componentwise relative perturbations may model the errors in some sparse matrix problems well but are not satisfactory for many other structured matrix problems.

Currently, no satisfactory computational method exists to evaluate the sensitivity of eigenvalue problems to perturbation classes other than norm-bounded perturbations, except in special cases. This includes the sensitivity of a simple eigenvalue and corresponding eigenvector of a general matrix to componentwise relative perturbations [3], [5], [17] and the sensitivity of all eigenvalues and eigenvectors of a symmetric positive definite matrix to the same [10], [15], [14]. The former results are too limited to be of much practical use and, although the latter results are very useful where they apply, they have not been extended to general matrices.

The theory of small-sample statistical condition estimation, recently introduced in [31], provides a tool to estimate the local sensitivity of any differentiable function at a point under a wide variety of assumptions on the type of perturbations of the point. In particular, it provides a way to monitor the effects of componentwise or structured perturbations to nonsingular linear systems of equations [33] and linear least-squares problems [32]. This paper proposes using the same theory to evaluate the sensitivity of eigenvalue problems and outlines how this can be accomplished for a general matrix and for a very large class of perturbations, including both componentwise relative and structured perturbations. Thus, it fills a gap in the set of computational resources available for the eigenvalue problem. For nonsymmetric matrices, the computational effort required is in most cases comparable with that currently accepted for norm-bounded perturbations [1] and is at most on the order of the effort required for computing the eigenvalues or invariant subspaces. For symmetric matrices, some specialized error bounds are more efficient to compute, but the approach proposed in this paper provides a way to handle a much larger class of perturbations than any other existing method.

Remark. The idea of condition estimation by statistical methods is by no means new and has been used extensively, usually in the framework of Monte Carlo trials [23], [38]. It has also been proposed in a similar setting as in [31], with applications to linear systems, least squares problems, and eigenvalue problems [11], [16], [43], but this seems to be the first approach that is of practical use. Different applications of statistical methods in relation to condition include the analysis of the statistical properties of random matrices [12], [13], analysis of the pseudospectra of matrices [45], and certain approaches to qualitative computing [6].

Section 2 of this paper gives a brief overview over perturbation theory for eigenvalues and invariant subspaces, section 3 reviews the theory of statistical condition estimation, and section 4 applies this estimation method to the eigenvalue problem under different assumptions on the perturbations. Section 5 gives an algorithmic description of the resulting estimators, along with a few numerical examples, and some conclusions are made in section 6.

In what follows, the 2-norm $\|\cdot\|_2$ is used exclusively for vectors and the Frobenius norm $\|\cdot\|_F$ exclusively for matrices. The range of a matrix is denoted by $\mathcal{R}(\cdot)$ and its spectrum by $\Lambda(\cdot)$. The Kronecker product of two matrices is denoted by the operator $\otimes$, the vec $(\cdot)$ operator stacks the columns of a matrix into a vector, and the unvec $(\cdot)$ operator reverses this operation (with matrix size determined by context).

2. Perturbations of eigenvalues and invariant subspaces. Let a matrix $A \in \mathbb{R}^{n\times n}$ be perturbed by a matrix E, about which nothing is known except that it belongs to some set $\mathcal{E}$. How much can the eigenvalues and invariant subspaces of A change by such a perturbation?

As discussed in the introduction, a wealth of results is available on this problem when $\mathcal{E}$ is a set of norm-bounded perturbations, $\mathcal{E} = \{E : \|E\| \le \epsilon \|A\|\}$ for some norm and $\epsilon > 0$, but similar results for other perturbation classes have only been derived for certain special cases. In contrast, the results presented in this paper can be applied to a general matrix and, moreover, can be implemented for a very wide class of perturbations. These results are based on classical first-order perturbation results for eigenvalues and invariant subspaces of general matrices as presented in, e.g., [40].

Accordingly, assume that the matrix A has the block-Schur decomposition

$$U^H A U = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \tag{1}$$

where $U = [U_1\ U_2]$ is unitary, U_1 has k columns, and the spectrum of the $k \times k$ submatrix $A_{11} = U_1^H A U_1$ is separated from the spectrum of A_{22}

$$\Lambda(A_{11}) \cap \Lambda(A_{22}) = \emptyset. \tag{2}$$

Further, assume that A is perturbed by a random matrix E, drawn from a distribution $\mathcal{E}$, denote the perturbed matrix by $\hat{A} = A + E$, and partition it conformably with A as

$$\begin{aligned} U^H \hat{A} U &= \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{bmatrix} \\ &= \begin{bmatrix} A_{11} + E_{11} & A_{12} + E_{12} \\ E_{21} & A_{22} + E_{22} \end{bmatrix}. \end{aligned}$$

The question posed at the beginning of the section may now be rephrased as follows: for $E \in \mathcal{E}$, by how much can the spectrum of A_{11} and the range of U_1 differ from the corresponding subset of the eigenvalues and corresponding invariant subspace of $\hat{A}$? The following theorem is a slight modification of Theorem 4.1 in [40] and provides the necessary elements to answer this.

THEOREM 2.1. *Let $B \in \mathbb{R}^{n\times n}$ and the unitary $W = [W_1\ W_2]$ be such that*

$$W^H B W = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

where the right-hand side is partitioned conformably with W. Define the operator $\mathcal{T}_B$ by $\mathcal{T}_B(Q) = QB_{11} - B_{22}Q$. If the Riccati equation

$$B_{21} + B_{22}P - PB_{11} - PB_{12}P = 0 \tag{3}$$

has a solution P, then the columns of

$$\hat{W}_1 = (W_1 + W_2P)(I + P^HP)^{-1/2}$$

are orthonormal and span an invariant subspace of B, and the matrix

$$\hat{B}_{11} = \hat{W}_1^H B \hat{W}_1$$

is similar to the matrix $B_{11} + B_{12}P$.

Equation (3) *has a solution if $\mathcal{T}_B$ is invertible and $\|B_{21}\|_F\|B_{12}\|_F\|\mathcal{T}_B^{-1}\|_F^2 < \frac{1}{4}$. In this case, the solution satisfies*

$$\|P\|_F \leq 2\|B_{21}\|_F\|\mathcal{T}_B^{-1}\|_F.$$

The application of this result with B replaced by $\hat{A} = A + E$ and W replaced by U shows that the invariant subspace spanned by the columns of U_1 is perturbed to the span of the columns of $U_1 + U_2X$,

$$\mathcal{R}\{\hat{U}_1\} = \mathcal{R}\{U_1 + U_2X\},$$

where X solves the equation

$$E_{21} - X\hat{A}_{12}X = \mathcal{T}_{\hat{A}}(X). \tag{4}$$

If E is small compared with A and the spectrum of A_{11} is well separated from the spectrum of A_{22}, then

$$\|\mathcal{T}_{\hat{A}}^{-1}\|_F \approx \|\mathcal{T}_A^{-1}\|_F$$

is moderate; see, for example, section 4 of [40]. Under these assumptions, the theorem shows that $\|X\|_F$ is on the order of $\|E\|_F$. Then a first-order approximation of equation (4) is

$$\mathcal{T}_A(X) \approx E_{21},$$

where $E_{21} = U_2^HEU_1$.

The solution to this linear equation can be used as a measure of the error in the invariant subspace U_1 induced by the perturbation E. This first-order error approximation is denoted by Δ_{U_1},

$$\Delta_{U_1} = \mathcal{T}_A^{-1}(U_2^HEU_1). \tag{5}$$

Remark. The k subspace angles θ_i, $i = 1, 2, \ldots, k$, between $\mathcal{R}\{U_1\}$ and $\mathcal{R}\{\hat{U}_1\}$ are given by

$$\tan(\theta_i(\mathcal{R}\{U_1\}, \mathcal{R}\{\hat{U}_1\})) = \sigma_i(X),$$

where σ_i are the singular values of X, suitably ordered [8]. Thus X represents the rotation of the invariant subspace by the perturbation.

Theorem 2.1 may be used directly to obtain an expression for the spectrum of the perturbed upper block $\hat{U}_1^H \hat{A} \hat{U}_1$, but a more useful expression may be obtained by first block-diagonalizing A. Thus, define

$$T = \begin{bmatrix} I & Y \\ 0 & I \end{bmatrix},$$

where $A_{12} = YA_{22} - A_{11}Y$, giving

$$T^{-1}U^H AUT = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}.$$

Then define

$$T^{-1}U^H EUT = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}.$$

By Theorem 2.1, there is a matrix $\hat{V}_1$ with orthonormal columns such that the spectrum of

$$\hat{V}_1^H T^{-1} U^H (A+E) UT \hat{V}_1 = \hat{A}_{11}$$

is equal to the spectrum of $A_{11} + F_{11} + F_{12}P$, where P solves a Riccati equation. For small perturbations, it may be assumed that P is small enough to disregard the term $F_{12}P$, so

$$\Lambda(\hat{A}_{11}) \approx \Lambda(A_{11} + F_{11})$$

or

$$\Lambda(\hat{A}_{11}) \approx \Lambda(A_{11} + (U_1^H - YU_2^H)EU_1). \tag{6}$$

The eigenvalues of A_{11} may be closely clustered or degenerate, and their individual sensitivity thus not well defined. The sensitivity of the average eigenvalue is well defined, however, as long as the spectra of A_{11} and A_{22} are well separated [29], [48]. Therefore, a condition estimator for this average, denoted by

$$\mu(A_{11}) = \frac{\operatorname{trace}(A_{11})}{k}$$

is derived in the sequel. By (6), this average eigenvalue is perturbed to

$$\begin{aligned} \mu(\hat{A}_{11}) &\approx \mu(A_{11} + (U_1^H - YU_2^H)EU_1) \\ &= \mu(A_{11}) + \mu((U_1^H - YU_2^H)EU_1). \end{aligned}$$

The error in the average eigenvalue of A_{11} induced by the perturbation E can be measured by this first-order approximation

$$\Delta_\mu = \mu((U_1^H - YU_2^H)EU_1). \tag{7}$$

Remark. The issue of how small E must be for the approximations above to be valid is not addressed in this paper. Discussion of this may be found in [2], for example, where various global bounds are also given.

When an appropriate block-Schur decomposition of A is available, the errors induced by a perturbation matrix E can be computed via their definitions, (5) and (7), but this need not be done in practice. Rather, the direct first-order difference approximations

$$\|\Delta_{U_1}\|_F^2 \approx \sum_{i=1}^{k} \tan^2(\theta_i(\mathcal{R}\{U_1\}, \mathcal{R}\{\hat{U}_1\})) \tag{8}$$

and

$$\Delta_\mu \approx \mu(\hat{A}_{11}) - \mu(A_{11}) \tag{9}$$

can be used, where the right-hand sides are computed directly from the eigenvalues and invariant subspaces of A and $A + E$, respectively. These quantities may be obtained via any computational method deemed appropriate, including, but not limited to, a Schur decomposition.

3. Small-sample statistical condition estimation. Common sense and experience suggest that a numerical computation is sensitive if small changes in the input data produce large changes in the output. Usually, slightly different information is desired: what is the largest change in the output that can occur under some given assumptions on the changes in the input? In [31] these ideas are connected by showing how the change in the output resulting from a small random perturbation of the input, *when properly scaled*, provides an estimate of the largest possible change. This technique is practical because the scaling factor between random and maximal change depends only on the number of input variables and is independent of the function whose value is being computed. Moreover, the maximal sensitivity can usually be adequately estimated by looking at the effect of very few random perturbations, and in many cases just one perturbation gives sufficient information. In this section a brief review of this theory of statistical condition estimation is given. First, an estimator for the sensitivity of scalar-valued functions of several variables is defined and its statistical properties described, and then this is extended to vector-valued functions.

Let the function $f : \mathbb{R}^p \to \mathbb{R}$ be differentiable at a point $x \in \mathbb{R}^p$ and assume that x is perturbed by a vector δz, where $\|z\|_2 \leq 1$ and $\delta > 0$ is small. Then

$$f(x + \delta z) \approx f(x) + \delta f_x(x) z,$$

where $f_x(x)$ is the Fréchet derivative of f at x

$$f_x(x) = \frac{\partial f}{\partial x}(x).$$

The relative error in f induced by the perturbation δz is given by

$$\frac{f(x + \delta z) - f(x)}{\delta} \approx f_x(x) z, \tag{10}$$

and the largest such error is given by $\|f_x(x)\|$. This latter quantity is a standard measure of the sensitivity, or condition of f at x [30], [37].

The derivative $f_x(x)$ is usually expensive to compute, but the product $f_x(x)z$ can often be evaluated efficiently for any z, either by using a first-order perturbation

analysis or by employing (10) directly. Thus, a method for estimating the norm of a vector given its inner product with another vector applies naturally to the problem of estimating the condition of f. The results of [31] are based on this approach and are presented in the framework of estimating vector norms.

If f is a vector-valued function, $f : \mathbb{R}^p \to \mathbb{R}^q$, the derivative f_x at a point x is a $q \times p$ matrix, rather than a row vector. As in the scalar case, a first-order perturbation analysis or the approximation (10) can be used efficiently to approximate the product of f_x and a vector, and a method for estimating matrix norms, given such products, applies to estimating the condition of f. A method for performing this estimation is presented in [21].

For the eigenvalue problem, x represents the elements of a matrix A, δz represents a perturbation E, and f and $\|f_x(x)\|$ represent the average eigenvalue and its sensitivity, or the subspace rotation and the corresponding sensitivity, respectively. Different definitions of the perturbation class $\mathcal{E}$ lead to different precise definitions of these quantities, as shown in the following section, but equations (5) and (7) can be used to compute $f_x(x)z$ in practice, or the product can be approximated by evaluating the function at x and $x + \delta z$ as in (8) and (9).

The distribution underlying the theory of statistical condition estimation is the uniform distribution of m-dimensional subspaces of $\mathbb{R}^p$, denoted by Ω_p^m. A basis for a sample from this distribution can be generated by selecting m independent vectors from the p-dimensional standard normal distribution $\mathcal{N}_p(0, 1)$ [28]. This basis is assumed to be orthonormalized. The results are presented in terms of the Wallis factor ω_p for an integer p

$$\omega_p = \begin{cases} 1 & \text{for } p = 1, \\ \frac{2}{\pi} & \text{for } p = 2, \\ \frac{1 \cdot 3 \cdot 5 \cdots (p-2)}{2 \cdot 4 \cdot 6 \cdots (p-1)} & \text{for odd } p > 2, \\ \frac{2}{\pi} \frac{2 \cdot 4 \cdot 6 \cdots (p-2)}{3 \cdot 5 \cdot 7 \cdots (p-1)} & \text{for even } p > 2. \end{cases}$$

This quantity can be approximated accurately, even for moderate values of p, by

$$\omega_p \approx \sqrt{\frac{2}{\pi(p - \frac{1}{2})}}.$$

THEOREM 3.1 (see [31]). *Let the columns of $Z \in \mathbb{R}^{p \times m}$ be an orthonormal basis for $\mathcal{Z} \sim \Omega_p^m$, and define*

$$\phi = \frac{\omega_m}{\omega_p} \|\ell^T Z\|_F$$

for $\ell \in \mathbb{R}^p$. Then $\mathrm{E}(\phi) = \|\ell\|_2$ and, for $\gamma > 1$,

$$\Pr\left(\frac{\|\ell\|_2}{\gamma} \le \phi \le \gamma \|\ell\|_2 \right) = g_{p,m}\left(\frac{\gamma \omega_p}{\omega_m} \right) - g_{p,m}\left(\frac{\omega_p}{\gamma \omega_m} \right)$$

where, for $0 \le \eta \le 1$, we have $g_{p,m}(\eta) = \frac{T_{p,m}(\eta)}{T_{p,m}(1)}$ with

$$T_{p,m}(\eta) = \int_0^\eta (1 - \nu^2)^{(p-m-2)/2} \nu^{m-1} \, d\nu.$$

Furthermore, the probability can be bounded by

$$\Pr\left(\frac{\|\ell\|_2}{\gamma} \le \phi \le \gamma\,\|\ell\|_2\right) \ge 1 - \frac{1}{m!}\left(\frac{2m}{\pi\gamma}\right)^m + \mathrm{O}\left(\frac{1}{\gamma^{m+1}}\right).$$

This result shows that for a small number of evaluations of the form $\ell^T z$, the probability of an estimate being within a small factor of the correct norm of ℓ is very high. For example, for $m = 3$ the probability of getting an estimate that is within a factor of 10 ($\gamma = 10$) of $\|\ell\|_2$ is greater than 0.9988, while for $\gamma = 100$, it is greater than 0.9999988.

In [21], the above is extended to matrix norms, rather than vector norms, thus allowing the same methods to be applied to functions into subspaces of order greater than one.

THEOREM 3.2 (see [21]). *Let the columns of $Z \in \mathbb{R}^{p\times m}$ be an orthonormal basis for $\mathcal{Z} \sim \Omega_p^m$, and define*

$$\psi = \sqrt{\frac{p}{m}}\|LZ\|_F$$

for $L \in \mathbb{R}^{q\times p}$. Then,

$$\mathrm{E}(\psi^2) = \|L\|_F^2.$$

Although not proved, the following is supported by extensive numerical evidence, and can almost certainly be taken to be true for the purpose of condition estimation.

CONJECTURE 3.3 (see [21]). *For the estimator ψ defined in Theorem 3.2 and for $\gamma > 1$,*

$$\Pr\left(\frac{\|L\|_F}{\gamma} \le \psi \le \gamma\,\|L\|_F\right) \ge g_{p,m}\left(\frac{\gamma\,\omega_p}{\omega_m}\right) - g_{p,m}\left(\frac{\omega_p}{\gamma\,\omega_m}\right),$$

where $g_{p,m}(y)$ is defined in Theorem 3.1.

Remark. When L is a row vector, $L = \ell^T$, the estimators ϕ and ψ do not coincide. Their statistical properties are effectively the same, however, in the sense that

$$\sqrt{\mathrm{E}(\psi^2)} = \mathrm{E}(\phi) = \|\ell\|_2$$

and

$$\Pr\left(\frac{\|\ell\|_2}{\gamma} \le \psi \le \gamma\,\|\ell\|_2\right) \ge \Pr\left(\frac{\|\ell\|_2}{\gamma} \le \phi \le \gamma\,\|\ell\|_2\right).$$

4. Condition estimation for eigenvalue problems. To facilitate appropriate definitions of condition numbers for different classes of perturbations, equations (5) and (7) can be rewritten.

LEMMA 4.1. *The first-order error in equation (7) can be written as*

$$\Delta_\mu = t^T \mathrm{vec}\,(E),$$

where

$$t = (U_1 \otimes (\bar{U}_1 - \bar{U}_2 Y^T))\frac{\mathrm{vec}\,(I_k)}{k}.$$

Similarly, equation (5)*, which is equivalent to solving the Sylvester equation* $XA_{11} - A_{22}X = U_2^H E U_1$*, can be written in the form*

$$\operatorname{vec}(\Delta_{U_1}) = R \operatorname{vec}(E),$$

where

$$R = (A_{11}^T \otimes I - I \otimes A_{22})^{-1}(U_1^T \otimes U_2^H).$$

Proof. This follows from applying the rules

$$\begin{aligned} \operatorname{vec}(K_1 K_2 K_3) &= (K_3^T \otimes K_1) \operatorname{vec}(K_2), \\ (K_1 \otimes K_2)^T &= K_1^T \otimes K_2^T, \\ \operatorname{trace}(K) &= (\operatorname{vec}(I))^T \operatorname{vec}(K) \end{aligned}$$

to the respective equations. ∎

For a general function f, the first-order approximation (10) suggests a definition of a condition number as $\|f_x(x)\|$. This gives the error in f relative to the size of the perturbation, indicating that all perturbations of equal norm are equally important. If the class of perturbations is such that this is not the case, namely, such that not all entries of the perturbation have equal weight, this needs to be modified. In fact, if the perturbation δz in the argument of f can be written as $\delta z = \delta S \tilde{z}$, where S is a constant matrix and $\|\tilde{z}\|$ is bounded, then $\|f_x(x)S\|$ can be taken as an appropriate condition number. Accordingly, condition numbers for the eigenvalue problem may be defined as follows.

DEFINITION 4.2. *Let the class of perturbations of* A *be given by*

$$\mathcal{E} = \{E : E = \delta \operatorname{unvec}(Sz),\ \delta > 0,\ z \in \mathbb{R}^p,\ \|z\|_2 \le 1\}$$

for some fixed matrix S*. Then a condition number for the average eigenvalue of the submatrix* A_{11} *is*

$$\tau(A, \mathcal{E}) = \|t^T S\|_2, \tag{11}$$

and a condition number for the corresponding invariant subspace is

$$\rho(A, \mathcal{E}) = \|RS\|_F, \tag{12}$$

where t *and* R *are defined in Lemma* 4.1*.*

Remark. The dimensions of the vector t and matrices R and S can be very high (for example, S can be as large as $n^2 \times n^2$), but they need not usually be computed explicitly in practice. In fact, the operations of R and t on any perturbation matrix are defined by (5) and (7) and approximated by (8) and (9), and S can usually be provided in the form of a set of rules for assigning the elements of the vector δz to the perturbation matrix E.

Estimators for the condition numbers $\tau(A, \mathcal{E})$ and $\rho(A, \mathcal{E})$ can now be defined.

DEFINITION 4.3. *Let* $Z \in \mathbb{R}^{p \times m}$ *be an orthonormal basis for* $\mathcal{Z} \sim \Omega_p^m$*, let* z_i*,* $i = 1, 2, \ldots, m$*, denote the columns of* Z*, and let* S *be such that* $\operatorname{unvec}(\delta S z_i)$ *is an element of* $\mathcal{E}$ *for each* $i = 1, 2, \ldots, m$*. Then estimators for the condition of the average of the spectrum of* A_{11} *and the corresponding invariant subspace are*

$$\phi_\tau = \frac{\omega_m}{\omega_p} \|t^T S Z\|_F \tag{13}$$

and

$$\psi_\rho = \sqrt{\frac{p}{m}}\,\|RSZ\|_F, \tag{14}$$

respectively.

The statistical properties of these estimators can be expressed as follows.

THEOREM 4.4. *For the estimator* ϕ_τ

$$\mathrm{E}(\phi_\tau) = \tau(A,\mathcal{E})$$

and

$$\Pr\left(\frac{\tau(A,\mathcal{E})}{\gamma} \le \phi_\tau \le \gamma\,\tau(A,\mathcal{E})\right) \ge 1 - \frac{1}{m!}\left(\frac{2m}{\pi\gamma}\right)^m + \mathrm{O}\left(\frac{1}{\gamma^{m+1}}\right)$$

for $\gamma > 1$. *Similarly, for the estimator* ψ_ρ

$$\mathrm{E}(\psi_\rho^2) = \rho^2(A,\mathcal{E})$$

and, under the hypothesis that Conjecture 3.3 *holds,*

$$\Pr\left(\frac{\rho(A,\mathcal{E})}{\gamma} \le \psi_\rho \le \gamma\,\rho(A,\mathcal{E})\right) \ge 1 - \frac{1}{m!}\left(\frac{2m}{\pi\gamma}\right)^m + \mathrm{O}\left(\frac{1}{\gamma^{m+1}}\right)$$

for $\gamma > 1$.

Proof. Apply Theorems 3.1 and 3.2 to the estimators as defined in Definition 4.3, noting the definitions of the condition numbers in (11) and (12). ∎

The following examples of different perturbation classes show the versatility of the estimators ϕ_τ and ψ_ρ.

4.1. Norm-bounded perturbations. The first case is that of norm-bounded perturbations

$$\|E\|_F \le \epsilon\,\|A\|_F$$

for a scalar $\epsilon > 0$. For this class of perturbations, other methods for estimating the condition of the eigenproblem are available [1], [10], some of which may be more efficient than the statistical approach presented here. This approach is, however, more widely applicable than those methods, since it neither puts restrictions on the matrix nor relies on a particular method for computing eigenelements.

The class of norm-bounded perturbations can be characterized as

$$\mathcal{E}_{nb} = \left\{E : E = \epsilon\,\|A\|_F z,\ z \in \mathbb{R}^{n^2},\ \|z\|_2 \le 1\right\}.$$

Referring to Definition 4.2, the matrix δS is trivially defined by

$$\delta S = \epsilon\,\|A\|_F\, I_{n^2},$$

and the condition estimates ϕ_τ and ψ_ρ are easily computed by evaluating (5) and (7) (or (8) and (9)) for m different perturbations of the form $E = \epsilon\,\|A\|_F\,\mathrm{unvec}\,(z)$ as specified in Definition 4.3. Note that the matrix δS is never explicitly needed in practice.

Remark. Applying the estimators ϕ_τ and ψ_ρ with $E \in \mathcal{E}_{nb}$ effectively gives estimates of the norm of the spectral projector for $\Lambda(A_{11})$ and the inverse of the separation $\mathrm{sep}\,(A_{11}, A_{22})$ between A_{11} and A_{22}, respectively, as those quantities are defined in, e.g., [44].

4.2. Componentwise relative perturbations. A perturbation class of special interest arises when the entries of A are perturbed relative to their size, or

$$|e_{ij}| \leq \epsilon\,|a_{ij}| \tag{15}$$

for some fixed scalar ϵ and $i, j = 1, 2, \ldots, n$. As discussed in the introduction, this is often considered the most appropriate description of certain types of perturbations such as rounding errors. Examples 2 and 3 demonstrate this. Note that the number of perturbation entries in E is equal to the number of nonzero entries of A, since the zero entries are not perturbed.

The class of componentwise relative perturbations can be characterized as follows. Let p denote the number of nonzero elements of A, let $a = [a_1\ a_2\ \cdots\ a_p]^T$ denote a vector consisting of those elements, let $C \in \mathbb{R}^{p\times n^2}$ be a matrix of ones and zeros such that $CC^T = I$ and $C^T a = \operatorname{vec}(A)$, and let

$$D = \begin{bmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_p \end{bmatrix}.$$

Then

$$\mathcal{E}_{cr} = \left\{E : E = \epsilon\, C^T D z,\ z \in \mathbb{R}^p,\ \|z\|_2 \leq 1\right\},$$

and the matrix δS in Definition 4.2 is given by

$$\delta S = \epsilon\, C^T D.$$

As for norm-bounded perturbations, this matrix need never be generated in practice, since an equivalent set of rules for assigning the elements of z to E is available. In fact, defining the matrix C is equivalent to assigning a unique integer $q_{ij} \in \{1, \ldots, p\}$ to the position of each nonzero entry a_{ij} of A, and the definition of δS is equivalent to setting $e_{ij} = \epsilon\, a_{ij}\, z_{q_{ij}}$ for each such entry. The remaining entries of E are zero.

The condition estimates ϕ_τ and ψ_ρ are easily computed by using these equivalent assignment rules to generate m perturbation matrices E as in Definition 4.3 and applying (5) and (7) (or (8) and (9)) to each of those.

4.3. Structured perturbations. Many different perturbation classes can arise when the matrix A has some fixed structure or shape. For example, some elements of A may be known to be equal, such as in symmetric matrices, some elements may be fixed, such as in sparse matrices (see, e.g., Example 1), or elements may depend on a single parameter, as in matrices arising from modeling some physical phenomenon. This section focuses on the case when certain elements of A are known to be equal and the perturbation is otherwise norm-bounded, $\|E\|_F \leq \epsilon\,\|A\|_F$. In this case, the number p of independent random parameters of the perturbation is equal to the number of independent elements of A.

The class of such structured perturbations can be characterized as follows. Let p be the number of independent elements of A and let $C \in \mathbb{R}^{p\times n^2}$ be such that $CC^T = I$ and the vector $C\operatorname{vec}(A)$ includes one and only one copy of each such element. Then

$$\mathcal{E}_s = \left\{E : E = \epsilon\,\|A\|_F\, C^T z,\ z \in \mathbb{R}^p,\ \|z\|_2 \leq 1\right\}.$$

In this case the matrix δS in Definition 4.2 is

$$\delta S = \epsilon \|A\|_F C^T,$$

which, in practice, may be replaced by an equivalent set of rules for assigning the elements of z to E. This set of rules states which entries of E should be equal and which entry of z should be assigned to each such group of perturbation entries. Using these rules to generate perturbations for equations (5) and (7) (or (8) and (9)) the condition estimators in Definition 4.3 can easily be computed.

Remark. Although only three examples of perturbation classes have been given here, any eigenvalue problem where the perturbations of the matrix elements can be modeled as in Definition 4.2 can be handled in the same fashion. The only requirement is that a matrix S can be provided, or, equivalently, a set of rules that describes a linear relationship between unit-norm vectors z that perturb the underlying parameters in an identical fashion and the corresponding effect on the elements of A. In particular, perturbations may be both structured and componentwise relative at the same time, as in the case of badly scaled symmetric matrices. The results of sections 4.2 and 4.3 can be combined to define appropriate assignment rules for this case.

5. Numerical examples. The results of the paper can be illustrated by restating them in the form of two pairs of algorithms. For the first pair, Algorithms 1 and 2, it is assumed that a block-Schur decomposition of a matrix $A \in \mathbb{R}^{n\times n}$ is given as (1) and that a linear relationship between the perturbations E and the random samples z is available, either in the form of a matrix or a set of assignment rules. Note that the perturbation matrices E have been scaled so that the factor ϵ can be omitted from these two algorithms.

ALGORITHM 1. *Given the decomposition* (1) *of* A*, the set of assignment rules* S*, the number of independent variables* p*, and the number of samples* m*, this algorithm estimates the condition of the average eigenvalue of* A_{11}*.*

Let the columns of Z *be an orthonormal basis for* $\mathcal{Z} \sim \Omega_p^m$*.*
Solve the equation $A_{12} = A_{11}Y - YA_{22}$ *for* Y*.*
For $i = 1, \ldots, m$
  Set z *equal to column* i *of* Z*.*
  Set $E = \text{unvec}\,(Sz)$*.*
  Form $E_{11} = U_1^H E U_1$ *and* $E_{21} = U_2^H E U_1$*.*
  Compute $\phi_i = |\mu(E_{11} + YE_{21})|$*.*
End for
Compute the condition estimate

$$\phi_\tau = \frac{\omega_m}{\omega_p}\sqrt{\sum_{i=1}^{m}\phi_i^2}.$$

ALGORITHM 2. *Given the decomposition* (1) *of* A*, the set of assignment rules* S*, the number of independent variables* p*, and the number of samples* m*, this algorithm estimates the condition of the invariant subspace spanned by the columns of* U_1*.*

Let the columns of Z *be an orthonormal basis for* $\mathcal{Z} \sim \Omega_p^m$*.*
For $i = 1, \ldots, m$
  Set z *equal to column* i *of* Z*.*
  Set $E = \text{unvec}\,(Sz)$*.*
  Solve the equation $E_{21} = XA_{11} - A_{22}X$ *for* X*.*

Compute $\psi_i = \|X\|_F$.
End for
Compute the condition estimate

$$\psi_\rho = \sqrt{\frac{p}{m}\sum_{i=1}^{m}\psi_i^2}.$$

The dominant calculation in each of the above algorithms is the solution of the appropriate Sylvester equation. The Golub–Nash–Van Loan algorithm [18] is especially well suited to this situation. In particular, for the Sylvester equation $AX + XB = C$ where $A \in \mathbb{R}^{\alpha\times\alpha}$ and $B \in \mathbb{R}^{\beta\times\beta}$ with $\alpha \geq \beta$, their algorithm takes approximately $\frac{5}{3}\alpha^3+10\beta^3+5\alpha^2\beta+\frac{5}{2}\alpha\beta^2$ floating-point operations for general A, B. However, if these coefficient matrices are already in Schur form, the operation count is only $3\alpha^2\beta+\frac{1}{2}\alpha\beta^2$. In Algorithms 1 and 2, the Sylvester equation coefficient matrices $A_{11} \in \mathbb{R}^{k\times k}$ and $A_{22} \in \mathbb{R}^{(n-k)\times(n-k)}$ are already in Schur form. Moreover, with $k \ll n$, the total operation count for the above algorithms, including both the Sylvester equation solution and associated matrix multiplications, is on the order of n^2km.

Apart from the factor m, this is the same order of magnitude as the LAPACK condition estimator for general eigenvalue problems [1]. Since $m \leq 3$ is usually sufficient (see further discussion below), the computational effort for our statistical approach is compatible with the conventional method, while applying to a wider class of perturbations. Condition estimators for some special problems, e.g., for componentwise relative perturbations of Hermitian matrices (see, e.g., [10]) may be more efficient than the statistical approach, and these should be used when they apply.

Each of the second pair of algorithms assumes that the average μ_0 of a set of eigenvalues of A has been computed along with a basis $\mathcal{U}_0$ for the corresponding invariant subspace. We also assume that a linear relationship S between the perturbations E and the random samples z is available.

ALGORITHM 3. *Given the average eigenvalue* μ_0*, the set of assignment rules* S*, the number of independent variables* p*, the number of samples* m*, and the size of the perturbation* ϵ*, this algorithm estimates the condition of* μ_0.

Let the columns of Z *be an orthonormal basis for* $\mathcal{Z} \sim \Omega_p^m$.
For $i = 1, \ldots, m$
Set z *equal to column* i *of* Z.
Set $E = \epsilon\,\mathrm{unvec}\,(Sz)$.
Compute the average μ_i *of a set of eigenvalues of* $A + E$ *corresponding to* μ_0.
Compute $\phi_i = |\mu_i - \mu_0|$.
End for
Compute the condition estimate

$$\phi_\tau = \frac{1}{\epsilon}\,\frac{\omega_m}{\omega_p}\sqrt{\sum_{i=1}^{m}\phi_i^2}.$$

Remark. Algorithm 3 can easily be modified for simultaneously estimating the condition of more than one eigenvalue or eigenvalue cluster by letting μ_0 denote a vector of the eigenvalues and average eigenvalues of interest, letting μ_i denote the corresponding vectors for the perturbed matrices, and computing ϕ_i and ϕ_τ for each entry of these vectors. In this case, clusters that need to be distinguished must

of course be sufficiently well separated to be uniquely identifiable for all allowable perturbations, but a similar modification of the algorithm also allows for estimation of the condition of multiple or closely clustered eigenvalues. In fact, the maximum perturbation of an eigenvalue in the cluster can be used as a measure of this condition, rather than the perturbation of the average eigenvalue in the cluster. Example 1 illustrates this.

ALGORITHM 4. *Given the basis $\mathcal{U}_0$ for an invariant subspace of A, the set of assignment rules S, the number of independent variables p, the number of samples m, and the size of the perturbation ϵ, this algorithm estimates the condition of $\mathcal{U}_0$.*

Let the columns of Z be an orthonormal basis for $\mathcal{Z} \sim \Omega_p^m$.

For $i = 1, \ldots, m$

Set z equal to column i of Z.

Set $E = \epsilon \operatorname{unvec}(Sz)$.

Compute a basis $\mathcal{U}_i$ for the invariant subspace of $A + E$ corresponding to $\mathcal{U}_0$.

Compute

$$\psi_i^2 = \sum_{j=1}^{k} \tan^2(\theta_j(\mathcal{U}_0, \mathcal{U}_i)).$$

End for

Compute the condition estimate

$$\psi_\rho = \frac{1}{\epsilon} \sqrt{\frac{p}{m} \sum_{i=1}^{m} \psi_i^2}.$$

The latter pair of algorithms requires on the order of m times the computational effort needed for computing μ_0 and $\mathcal{U}_0$, respectively, which can be significantly more than for the former pair of algorithms. Nevertheless, the wide class of perturbations to which they apply make them valuable alternatives to existing methods.

Choice of m. The condition estimates given below in Example 1 are the average over several applications of Algorithm 3 with $m = 1$. The probability of obtaining an estimate close to the mean value depends, of course, on the number of samples m as discussed in section 3. For example, the probability of obtaining an estimate within a factor of 10 of the mean is in theory approximately bounded below by

$$1 - \frac{1}{m!} \left(\frac{m}{5\pi} \right)^m,$$

which is close to unity, even for small m. Numerical results agree well with this theoretical bound, and in the remaining examples the value of m is left unspecified (we actually used $m = 1$).

Example 1. A standard example of an eigenvalue problem where bounds based on norm-bounded perturbations overestimate the actual perturbation by a large margin is an $n \times n$ matrix composed of a single Jordan block. It is well known [29] that a perturbation of size ϵ of the zero in the lower left-hand corner causes the eigenvalues to spread out on a circle of radius $\epsilon^{1/n}$ around the multiple eigenvalue of the unperturbed matrix. If the shape of the matrix is known, however, the zero elements may be assumed not to change by the perturbation, and the eigenvalues of the perturbed matrix are contained in a circle of radius ϵ around the original eigenvalue.

The small-sample statistical estimation approach predicts this behavior accurately. As a concrete example, let

$$A = \begin{bmatrix} 0.2 & 1 & & & & & \\ & 0.2 & 1 & & & & \\ & & & \ddots & & & \\ & & & & 0.2 & 1 & \\ & & & & & 0.2 \end{bmatrix}$$

be of order $n = 10$ and let $\epsilon \approx 2 \times 10^{-16}$. Since all the eigenvalues are equal, assumption (2) is not satisfied, and Algorithm 3 must be used to estimate their condition. If norm-bounded perturbations $\|E\|_F = \epsilon \|A\|_F$ are assumed, the algorithm gives error estimates of approximately $10^{15}\,\epsilon$ for the computed eigenvalues, which indicates that no digits can be assumed to be correct. If the zero elements of A are not perturbed, however, the same algorithm gives error estimates of approximately ϵ. Note that in the former case, $p = n^2$ and the assignment rule S given to the algorithm is simply that the ith entry of $\mathrm{vec}\,(E)$ equals the ith entry of z multiplied by $\epsilon \|A\|_F$. In the latter case, however, $p = 2n - 1$ equals the number of nonzero entries of A, and S assigns the p entries of each vector $\epsilon \|A\|_F z$ to the corresponding p entries of E and sets the remaining entries to zero. In both cases, $\mu_0 = 0.2$ and the condition is estimated for individual eigenvalues by monitoring the maximum perturbation of the spectrum of A, rather than the perturbation of the average eigenvalue.

Example 2. An example of how the assumption of componentwise relative perturbations may be more appropriate than full additive perturbations follows. Let

$$A_0 = Q - I_n,$$

where $q_{ij} = 1$ for $i, j = 1, 2, \ldots, n$, let

$$M = \begin{bmatrix} 1 & & & \\ & \xi & & \\ & & \ddots & \\ & & & \xi^{n-1} \end{bmatrix},$$

where ξ is a positive scalar, and let $A = MA_0M^{-1}$. This matrix is Toeplitz, has $n - 1$ eigenvalues at -1 and one at $n - 1$ and is badly scaled for ξ not close to 1. Additionally, the eigenvector corresponding to the eigenvalue at $n - 1$ is proportional to

$$x = \begin{bmatrix} 1 & \xi & \xi^2 & \cdots & \xi^{n-1} \end{bmatrix}^T.$$

For a Schur decomposition of A with $A_{11} = n - 1$ and all eigenvalues of A_{22} at -1, the upper bound on the subspace angle between the computed eigenvector $\hat{x}$ and the corresponding vector for $A + E$ for norm-bounded perturbations $\|E\|_F \leq \epsilon \|A\|_F$ is very high, while the decomposition is computed very accurately in practice. The reason for this is, of course, that the matrix is badly scaled and its entries are not perturbed equally by rounding errors. Therefore, componentwise relative perturbations are a more appropriate description of the errors than norm-bounded perturbations.

As a numerical example, take $\xi = 3$ and $n = 20$. An upper bound on the subspace angle induced by a perturbation matrix E as above is approximately

$$\frac{\|E\|_F}{\mathrm{sep}_F(A_{11}, A_{22})} \leq \frac{\epsilon \|A\|_F}{\mathrm{sep}_F(A_{11}, A_{22})} \approx 3 \times 10^8 \, \epsilon,$$

and applying Algorithm 2 to this case, with the assignment rules S for norm-bounded perturbations as described in section 4.1, gives a similar value. In contrast, when componentwise relative perturbations are assumed and the assignment rules S are constructed as described in section 4.2, the estimates given by Algorithm 2 are on the order of $10^{-1}\,\epsilon$. Note that for simplicity of exposition the Toeplitz structure of A is not exploited in this example. In fact, incorporating this structural information in the assignment rules S may give an even lower estimate of the condition of the problem.

The sensitivity of the simple eigenvalue at $n-1$ shows similar behavior, with the upper bound approximately $10^{17}\,\epsilon$ when norm-bounded perturbations are assumed in Algorithm 1, but on the order of ϵ when componentwise relative perturbations are assumed.

Example 3. The great discrepancy between the condition estimates in the previous example is primarily due to the matrix A being very badly scaled. Another such badly scaled example is given in [10]. In this case the matrix is symmetric and, unlike before, the QR method fails to compute a Schur form for the matrix. Therefore, Algorithms 3 and 4 must be employed, with Jacobi iterations used to compute the eigenvalues for each perturbation.

The matrix under consideration is

$$A = \begin{bmatrix} 10^{40} & 10^{29} & 10^{19} \\ 10^{29} & 10^{20} & 10^{9} \\ 10^{19} & 10^{9} & 1 \end{bmatrix}$$

with eigenvalues at approximately 10^{40}, 9.9×10^{19}, and 0.98. As discussed in [10], the QR method fails to compute the two smaller eigenvalues, but all eigenvalues are computed accurately via Jacobi iterations. According to the authors, the condition number of all the eigenvalues is on the order of 10^{40}, while componentwise relative error bounds for symmetric matrices give a value of approximately 1.33. Applying Algorithm 3 (modified either directly or modified according to the remark following Algorithm 3 to estimate the sensitivity of the individual eigenvalues) with norm-bounded perturbations as described in section 4.1 supports this result and gives absolute error estimates on the order of $10^{40}\,\epsilon$ for each eigenvalue. On the other hand, if both the symmetry of the matrix and the size of individual components are exploited in the construction of the assignment rules S, the estimates are on the order of $\epsilon\,\lambda_i$ for each eigenvalue λ_i. Constructing these assignment rules may be done by noting that A has only $p = 6$ independent elements, which may be taken to be its upper triangle. Thus, each perturbation z is of dimension 6, its elements are assigned to the 6 elements of the upper triangle of A, each one is multiplied by ϵ and the corresponding entry of A, and a symmetric perturbation E is created from the resulting values.

Example 4. In some applications, the computation of the eigenelements of a matrix is only an intermediate step toward some desired final result and the question of the relevance of the accuracy of this particular step arises. More precisely, does an ill-conditioned eigenvalue problem necessarily indicate ill conditioning in the final result, and, vice versa, when does a well-conditioned eigenvalue step guarantee a well-conditioned final result? This example illustrates how this may arise and motivates some further work in that direction.

Computing one type of optimal control law for a linear n-dimensional differential system

$$\dot{x}(t) = Fx(t) + Gu(t),$$
$$y(t) = Hx(t),$$

where t denotes time, $x(t) \in \mathbb{R}^n$, $u(t)$ and $y(t)$ are vectors or scalars, and F, G, and H are appropriately dimensioned matrices, involves solving a Riccati equation of the form [4]

$$F^T X + FX - XGG^T X + H^T H = 0.$$

When certain system-theoretic assumptions on F, G, and H are satisfied (see, for example, [4]), a unique symmetric positive definite solution to this equation exists, the so-called stabilizing solution, which may be computed via the $2n \times 2n$ Hamiltonian matrix

$$A = \begin{bmatrix} F & -GG^T \\ -H^T H & -F^T \end{bmatrix}.$$

This matrix has exactly n eigenvalues in the open left half-plane and, if the columns of $[X_1^T \ X_2^T]^T$ span the invariant subspace corresponding to those eigenvalues, then $X = X_2 X_1^{-1}$ is the desired stabilizing solution.

Obviously, both the condition of the invariant subspace computation and the condition of X_1 with respect to inversion contribute to the condition of the solution X. Moreover, the choice of X_1 is not unique and its condition may depend heavily on the choice of basis vectors for the invariant subspace [34]. Thus, the condition of this invariant subspace is not the only factor determining the condition of the Riccati solution, and to what extent it affects the final result is not obvious.

To illustrate, let

$$F = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -0.003 & 1 \\ 0 & 0 & -0.002 \end{bmatrix}, \quad G = \begin{bmatrix} 0.1 \\ 0 \\ 0 \end{bmatrix}, \quad H = I_3.$$

Assuming norm-bounded perturbations, the condition of the invariant subspace corresponding to the three eigenvalues of A in the left half-plane is on the order of 10^7. If the "shape" of F, G, and H is fixed and only the nonzero elements are assumed to be perturbed, the condition is only on the order of unity, showing that in this case the desired invariant subspace may be computed accurately. This holds for both componentwise relative perturbations and equal perturbations of the nonzero elements. For all three types of perturbations the condition of X_1 with respect to inversion is high, or $O(10^6)$.

The condition of the two steps does not seem to translate to the condition of the Riccati solution in a consistent fashion. In fact, the latter value is on the order of 10^7 for both norm-bounded perturbations and equal perturbations of the nonzero elements of F, G, and H but only on the order of unity for componentwise relative perturbations. This issue will be investigated further in a forthcoming paper.

6. Conclusions. A statistical approach to estimating the sensitivity of eigenvalue computations to perturbations in the matrix elements has been proposed. This approach usually requires a computational effort similar to that of standard condition

estimators for general matrices but more than some specialized ones. It has an added benefit, however, namely, that it allows for a much wider class of perturbations than other approaches, while giving results that have a high probability of being accurate. In particular, the class of matrix perturbations can be restricted to componentwise relative perturbations or various structured perturbations.

The statistical approach gives a powerful alternative to conventional condition estimators for eigenvalue problems. In addition to being applicable to a wide class of perturbations and requiring only moderate computational effort, it may be well suited to problems involving large matrices since only a subset of eigenvalues and the associated invariant subspace are required to evaluate the estimators. This is a subject for future investigation.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. D. Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, Lapack *Users' Guide*, 2nd ed., SIAM, Philadelphia, PA, 1994.

[2] Z. Bai, J. W. Demmel, and A. McKenney, *On computing condition numbers for the nonsymmetric eigenvalue problem*, ACM Trans. Math. Software, 19 (1993), pp. 202–223.

[3] F. L. Bauer and C. T. Fike, *Norms and exclusion theorems*, Numer. Math., 2 (1960), pp. 137–141.

[4] S. Bittanti, A. J. Laub, and J. C. Willems, eds., *The Riccati Equation*, Springer-Verlag, Berlin, 1991.

[5] F. Chatelin, *Eigenvalues of Matrices*, Wiley, Chichester, 1993.

[6] F. Chatelin and V. Frayssé, *Qualitative computing: Elements of a theory for finite precision computation*, Lecture notes, THOMPSON-CSF, Laboratoire Central de Recherches, Domaine de Corbeville, Orsay, France, June 1993.

[7] 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.

[8] C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation,* III, SIAM J. Numer. Anal., 7 (1970), pp. 1–46.

[9] J. W. Demmel and B. Kågström, *Computing stable eigendecompositions of matrix pencils*, Linear Algebra Appl., 88/89 (1987), pp. 139–186.

[10] J. W. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1246.

[11] J. D. Dixon, *Estimating extremal eigenvalues and condition numbers of a matrix*, SIAM J. Numer. Anal., 20 (1983), pp. 812–814.

[12] A. Edelman, *Eigenvalues and Condition Numbers of Random Matrices*, Ph.D. thesis, Massachusetts Institute of Technology, Dept. of Mathematics, Cambridge, MA, May 1989.

[13] A. Edelman, *The distribution and moments of the smallest eigenvalue of a random matrix of Wishart type*, Linear Algebra Appl., 159 (1991), pp. 55–80.

[14] S. C. Eisenstat and I. C. F. Ipsen, *Relative perturbation bounds for eigenspaces and singular vector subspaces*, in Proc. Fifth SIAM Conf. Applied Linear Algebra, J. Lewis, ed., Snowbird, UT, June 1994, pp. 62–65.

[15] S. C. Eisenstat and I. C. F. Ipsen, *Relative perturbation techniques for singular value problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1972–1988.

[16] R. Fletcher, *Expected conditioning*, IMA J. Numer. Anal., 5 (1985), pp. 247–273.

[17] A. J. Geurts, *A contribution to the theory of condition*, Numer. Math., 39 (1982), pp. 85–96.

[18] G. H. Golub, S. Nash, and C. F. Van Loan, *A Hessenberg-Schur method for the problem* $AX + XB = C$, IEEE Trans. Automat. Control, AC-24 (1979), pp. 909–913.

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

[20] G. H. Golub and J. H. Wilkinson, *Ill-conditioned eigensystems and the computation of the Jordan canonical form*, SIAM Rev., 18 (1976), pp. 578–619.

[21] T. Gudmundsson, C. S. Kenney, and A. J. Laub, *Small-sample statistical estimates for matrix norms*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 776–792.

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

[23] J. Hammersley and D. Handscomb, *Monte Carlo Methods*, Methuen and Wiley, London, 1964.

[24] 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.
[25] N. J. Higham, *A survey of componentwise perturbation theory in numerical linear algebra*, in Mathematics of Computation 1943-1993, Proc. Symposia in Applied Mathematics, Vancouver, BC, Canada, August 1993, W. Gautschi, ed., AMS, Providence, RI, pp. 49–77.
[26] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996.
[27] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, London, 1985.
[28] A. T. James, *Normal multivariate analysis and the orthogonal group*, Ann. Math. Statist., 25 (1954), pp. 40–75.
[29] W. Kahan, *Conserving Confluence Curbs Ill-Condition*, Computer Science Technical Report 6, University of California, Berkeley, CA, August 1972.
[30] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, Berlin, 1982.
[31] 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.
[32] C. Kenney, A. J. Laub, and M. S. Reese, *Statistical condition estimation for linear least-squares problems*, SIAM J. Matrix Anal. Appl., submitted.
[33] C. Kenney, A. J. Laub, and M. S. Reese, *Statistical condition estimation for linear systems*, SIAM J. Sci. Comput., to appear.
[34] C. S. Kenney, A. J. Laub, and M. Wette, *A stability-enhancing scaling procedure for Schur-Riccati solvers*, Systems Control Lett., 12 (1989), pp. 241–250.
[35] C. D. Meyer and G. W. Stewart, *Derivatives and perturbations of eigenvectors*, SIAM J. Numer. Anal., 25 (1988), pp. 679–691.
[36] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[37] J. R. Rice, *A theory of condition*, SIAM J. Numer. Anal., 3 (1966), pp. 287–310.
[38] H. Rief, *A synopsis of Monte Carlo perturbation algorithms*, J. Comput. Physics, 111 (1994), pp. 33–48.
[39] G. W. Stewart, *On the sensitivity of the eigenvalue problem* $Ax = \lambda Bx$, SIAM J. Numer. Anal., 9 (1972), pp. 669–686.
[40] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.
[41] G. W. Stewart, *On the perturbation of pseudo-inverses, projections and linear least squares problems*, SIAM Rev., 19 (1977), pp. 634–662.
[42] G. W. Stewart, *A second order perturbation expansion for small singular values*, Linear Algebra Appl., 56 (1984), pp. 231–235.
[43] G. W. Stewart, *Stochastic perturbation theory*, SIAM Rev., 32 (1990), pp. 579–610.
[44] G. W.Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, San Diego, CA, 1990.
[45] L. N. Trefethen, *Pseudo-spectra of matrices*, in Numerical Analysis 1991, D. Griffith and G. Watson, eds., Longman, Essex, UK, 1992.
[46] P.-A. Wedin, *Perturbation bounds in connection with the singular value decomposition*, BIT, 12 (1972), pp. 99–111.
[47] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.
[48] J. H. Wilkinson, *Sensitivity of eigenvalues*, Utilitas Mathematica, 25 (1984), pp. 5–76.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 4, pp. 887–892, October 1997

007

A NOTE ON A PARTIAL ORDERING IN THE SET OF HERMITIAN MATRICES*

JÜRGEN GROSS†

Abstract. In this note we introduce a partial ordering in the set of complex Hermitian matrices which coincides with the well-known Löwner ordering when the considered matrices have the same number of negative eigenvalues. Some properties of the new ordering are investigated, and known results for shorted matrices are generalized.

Key words. Hermitian matrix, partial ordering, shorted matrix

AMS subject classification. 15A57

PII. S0895479896305684

1. Introduction.

Let $\mathbb{C}_{m,n}$ denote the set of complex $m \times n$ matrices and $\mathbb{C}_m$ denote the set $\mathbb{C}_{m,m}$. Let $\mathbb{C}_m^H$ denote the subset of $\mathbb{C}_m$ containing all Hermitian matrices, and let $\mathbb{C}_m^{\geq}$ denote the subset of $\mathbb{C}_m^H$ containing all non-negative definite matrices. The symbols $\mathbf{A}^*$, $\mathcal{R}(\mathbf{A})$, $\mathcal{R}^{\perp}(\mathbf{A})$, $\mathcal{N}(\mathbf{A})$, and $\mathrm{rk}(\mathbf{A})$ will stand for the conjugate transpose, the range, the orthogonal complement of the range, the null space, and the rank, respectively, of $\mathbf{A} \in \mathbb{C}_{m,n}$. By $\mathbf{A}^+$ we denote the Moore–Penrose inverse of $\mathbf{A} \in \mathbb{C}_{m,n}$, i.e., the unique matrix $\mathbf{A}^+$ satisfying $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$, $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$, $(\mathbf{A}\mathbf{A}^+)^* = \mathbf{A}\mathbf{A}^+$, and $(\mathbf{A}^+\mathbf{A})^* = \mathbf{A}^+\mathbf{A}$. Any matrix $\mathbf{A}^-$ satisfying only the first of these equations, i.e., $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$, is called a generalized inverse of $\mathbf{A} \in \mathbb{C}_{m,n}$.

In this paper we are concerned with two known partial orderings. One of them is defined in the set of complex rectangular matrices, whereas the other is only applicable to square matrices. The rank subtractivity partial ordering $\mathbf{A} \overset{\mathrm{rs}}{\leq} \mathbf{B}$ in $\mathbb{C}_{m,n}$ is defined by

$$\mathbf{A} \overset{\mathrm{rs}}{\leq} \mathbf{B} \quad :\Leftrightarrow \quad \mathrm{rk}(\mathbf{B} - \mathbf{A}) = \mathrm{rk}(\mathbf{B}) - \mathrm{rk}(\mathbf{A}), \tag{1}$$

whereas the Löwner partial ordering $\mathbf{A} \overset{\mathrm{L}}{\leq} \mathbf{B}$ in $\mathbb{C}_m$ is defined by

$$\mathbf{A} \overset{\mathrm{L}}{\leq} \mathbf{B} \quad :\Leftrightarrow \quad \mathbf{B} - \mathbf{A} = \mathbf{K}\mathbf{K}^* \tag{2}$$

for some matrix $\mathbf{K}$. The right-hand part of this definition means that $\mathbf{B} - \mathbf{A}$ is a non-negative definite matrix, i.e., $\mathbf{B} - \mathbf{A} \in \mathbb{C}_m^{\geq}$.

When $\mathbf{A}$ and $\mathbf{B}$ are Hermitian matrices, then

$$\mathbf{A} \overset{\mathrm{rs}}{\leq} \mathbf{B} \quad \Leftrightarrow \quad \mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{B}) \text{ and } \mathbf{A}\mathbf{B}^+\mathbf{A} = \mathbf{A}; \tag{3}$$

cf. statement (1.21) in Baksalary, Pukelsheim, and Styan (1989). Moreover, when $\mathbf{A}$ and $\mathbf{B}$ are Hermitian matrices with the same number of negative eigenvalues, then

$$\mathbf{A} \overset{\mathrm{L}}{\leq} \mathbf{B} \quad \Leftrightarrow \quad \mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{B}) \text{ and } \mathbf{A}\mathbf{B}^+\mathbf{A} \overset{\mathrm{L}}{\leq} \mathbf{A}; \tag{4}$$

*Received by the editors June 24, 1996; accepted for publication by G. P. Styan September 23, 1996.

http://www.siam.org/journals/simax/18-4/30568.html

†Department of Statistics, University of Dortmund, Vogelpothsweg 87, D-44221 Dortmund, Germany (gross@amadeus.statistik.uni-dortmund.de). This work was supported by the Deutsche Forschungsgemeinschaft under grant Tr 253/2-1/2-2.

cf. Gross (1997). Observe that (4) has been established earlier for non-negative definite matrices $\mathbf{A}$ and $\mathbf{B}$; cf. statement (1.20) in Baksalary, Pukelsheim, and Styan (1989). In this note we show that the conditions

$$\mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{B}) \text{ and } \mathbf{A}\mathbf{B}^+\mathbf{A} \overset{\text{L}}{\leq} \mathbf{A}$$

specify themselves a partial ordering within the set of Hermitian matrices. By regarding the shorted matrix with respect to the new ordering, a generalization of the results of Goller (1986) and Mitra, Puntanen, and Styan (1994) is derived.

2. Ordering of Hermitian matrices. In the following we write $\mathbf{A} \overset{\circ}{\leq} \mathbf{B}$ whenever the Hermitian matrices $\mathbf{A}$ and $\mathbf{B}$ satisfy the conditions $\mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{B})$ and $\mathbf{A}\mathbf{B}^+\mathbf{A} \overset{\text{L}}{\leq} \mathbf{A}$. Note that in view of $\mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{B})$ the product $\mathbf{A}\mathbf{B}^+\mathbf{A}$ does not depend on the choice of generalized inverse; i.e., $\mathbf{A}\mathbf{B}^+\mathbf{A} = \mathbf{A}\mathbf{B}^-\mathbf{A}$ for every generalized inverse $\mathbf{A}^-$ of $\mathbf{A}$.

The following theorem ensures that the relation $\overset{\circ}{\leq}$ specifies, in fact, a partial ordering in the set of Hermitian matrices.

THEOREM 1. *For two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{C}_m^H$, the relation $\overset{\circ}{\leq}$ defined by*

$$\mathbf{A} \overset{\circ}{\leq} \mathbf{B} \quad :\Leftrightarrow \quad \mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{B}) \text{ and } \mathbf{A}\mathbf{B}^+\mathbf{A} \overset{\text{L}}{\leq} \mathbf{A}$$

specifies a partial ordering in $\mathbb{C}_m^H$.

Proof. Reflexivity of the relation $\overset{\circ}{\leq}$ is obvious.

Transitivity: Let $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{C}_m^H$ such that $\mathbf{A} \overset{\circ}{\leq} \mathbf{B}$ and $\mathbf{B} \overset{\circ}{\leq} \mathbf{C}$. Then $\mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{B})$ and $\mathcal{R}(\mathbf{B}) \subseteq \mathcal{R}(\mathbf{C})$, which clearly gives $\mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{C})$. Moreover, we have $\mathbf{A}\mathbf{B}^+\mathbf{A} \overset{\text{L}}{\leq} \mathbf{A}$, i.e., $\mathbf{A}(\mathbf{A}^+ - \mathbf{B}^+)\mathbf{A} \in \mathbb{C}_m^{\geq}$, and $\mathbf{B}\mathbf{C}^+\mathbf{B} \overset{\text{L}}{\leq} \mathbf{B}$, i.e., $\mathbf{B}(\mathbf{B}^+ - \mathbf{C}^+)\mathbf{B} \in \mathbb{C}_m^{\geq}$. Multiplying $\mathbf{B}(\mathbf{B}^+ - \mathbf{C}^+)\mathbf{B}$ from the left with $\mathbf{A}\mathbf{B}^+$ and from the right with $(\mathbf{A}\mathbf{B}^+)^* = \mathbf{B}^+\mathbf{A}$ shows that also $\mathbf{A}\mathbf{B}^+\mathbf{B}(\mathbf{B}^+ - \mathbf{C}^+)\mathbf{B}\mathbf{B}^+\mathbf{A} \in \mathbb{C}_m^{\geq}$. But in view of $\mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{B})$ we have $\mathbf{A}\mathbf{B}^+\mathbf{B} = \mathbf{A} = \mathbf{B}\mathbf{B}^+\mathbf{A}$, which gives $\mathbf{A}(\mathbf{B}^+ - \mathbf{C}^+)\mathbf{A} \in \mathbb{C}_m^{\geq}$ and thus $\mathbf{A}\mathbf{C}^+\mathbf{A} \overset{\text{L}}{\leq} \mathbf{A}\mathbf{B}^+\mathbf{A} \overset{\text{L}}{\leq} \mathbf{A}$. Hence $\mathbf{A} \overset{\circ}{\leq} \mathbf{C}$.

Antisymmetry: Let $\mathbf{A}, \mathbf{B} \in \mathbb{C}_m^H$ such that $\mathbf{A} \overset{\circ}{\leq} \mathbf{B}$ and $\mathbf{B} \overset{\circ}{\leq} \mathbf{A}$. Then $\mathcal{R}(\mathbf{A}) = \mathcal{R}(\mathbf{B})$. Moreover, $\mathbf{A}\mathbf{B}^+\mathbf{A} \overset{\text{L}}{\leq} \mathbf{A}$, i.e., $\mathbf{A}(\mathbf{A}^+ - \mathbf{B}^+)\mathbf{A} \in \mathbb{C}_m^{\geq}$, and $\mathbf{B}\mathbf{A}^+\mathbf{B} \overset{\text{L}}{\leq} \mathbf{B}$, i.e., $\mathbf{B}(\mathbf{B}^+ - \mathbf{A}^+)\mathbf{B} \in \mathbb{C}_m^{\geq}$. Multiplying $\mathbf{B}(\mathbf{B}^+ - \mathbf{A}^+)\mathbf{B}$ from the left with $\mathbf{A}\mathbf{B}^+$ and from the right with $(\mathbf{A}\mathbf{B}^+)^* = \mathbf{B}^+\mathbf{A}$ gives $\mathbf{A}(\mathbf{B}^+ - \mathbf{A}^+)\mathbf{A} \in \mathbb{C}_m^{\geq}$. But $\mathbf{A}(\mathbf{A}^+ - \mathbf{B}^+)\mathbf{A} \in \mathbb{C}_m^{\geq}$ and $\mathbf{A}(\mathbf{B}^+ - \mathbf{A}^+)\mathbf{A} \in \mathbb{C}_m^{\geq}$ can only hold together when $\mathbf{A}(\mathbf{B}^+ - \mathbf{A}^+)\mathbf{A} = \mathbf{0}$, i.e., $\mathbf{A}\mathbf{B}^+\mathbf{A} = \mathbf{A}$. Now, since we also have $\mathcal{R}(\mathbf{B}) \subseteq \mathcal{R}(\mathbf{A})$ there exists a matrix $\mathbf{G}$ such that $\mathbf{B} = \mathbf{A}\mathbf{G}$. Multiplying $\mathbf{A}\mathbf{B}^+\mathbf{A} = \mathbf{A}$ from the right with $\mathbf{G}$ gives $\mathbf{A}\mathbf{B}^+\mathbf{B} = \mathbf{B}$; i.e., $\mathbf{A} = \mathbf{B}$ in view of $\mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{B})$. □

Let $\nu(\mathbf{A})$ denote the number of negative eigenvalues of a matrix $\mathbf{A} \in \mathbb{C}_m^H$. Then we may state the following.

COROLLARY 1. *Let $\mathbf{A}, \mathbf{B} \in \mathbb{C}_m^H$. Then $\mathbf{A} \overset{\circ}{\leq} \mathbf{B}$ if and only if $\mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{B})$ and $\nu(\mathbf{B} - \mathbf{A}) = \nu(\mathbf{B}) - \nu(\mathbf{A})$.*

Proof. From statement (1.66) in Styan (1985) or Lemma 2 in Gross (1997), we get $\nu(\mathbf{A} - \mathbf{A}\mathbf{B}^+\mathbf{A}) = \nu(\mathbf{B} - \mathbf{A}) - [\nu(\mathbf{B}) - \nu(\mathbf{A})]$ when $\mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{B})$. Since $\mathbf{A}\mathbf{B}^+\mathbf{A} \overset{\text{L}}{\leq} \mathbf{A}$ if and only if $\nu(\mathbf{A} - \mathbf{A}\mathbf{B}^+\mathbf{A}) = 0$, the assertion follows immediately. □

Our next corollary has already been mentioned in the introduction. The nontrivial part of its proof appears to be the implication $\mathbf{A} \overset{\text{L}}{\leq} \mathbf{B} \Rightarrow \mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{B})$; cf. Gross (1997, Theorem).

COROLLARY 2. *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{C}_m^H$ *be such that* $\nu(\mathbf{A}) = \nu(\mathbf{B})$. *Then* $\mathbf{A} \overset{\circ}{\leq} \mathbf{B}$ *if and only if* $\mathbf{A} \overset{\text{L}}{\leq} \mathbf{B}$.

Observe that the ordering $\overset{\circ}{\leq}$ does not coincide with the Löwner partial ordering in general, even when $\mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{B})$. Choose, e.g., $\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Then $\mathbf{A} \overset{\text{L}}{\leq} \mathbf{B}$, but $\mathbf{A} \overset{\circ}{\leq} \mathbf{B}$ does not hold. On the other hand, by choosing $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 2 & 2 \\ 2 & -2 \end{pmatrix}$ we have $\mathbf{A} \overset{\circ}{\leq} \mathbf{B}$, but $\mathbf{A}$ and $\mathbf{B}$ are not ordered with respect to $\overset{\text{L}}{\leq}$. Note that the last two example matrices also confirm the fact that $\mathbf{A} \overset{\circ}{\leq} \mathbf{B}$ does not imply $\mathbf{A} \overset{\text{rs}}{\leq} \mathbf{B}$. Of course, the converse is true for arbitrary Hermitian matrices $\mathbf{A}$ and $\mathbf{B}$, i.e., $\mathbf{A} \overset{\text{rs}}{\leq} \mathbf{B} \Rightarrow \mathbf{A} \overset{\circ}{\leq} \mathbf{B}$.

Moreover, it is easily verified that the condition $\mathbf{A}\mathbf{B}^+\mathbf{A} \overset{\text{L}}{\leq} \mathbf{A}$ alone cannot specify a partial ordering. Consider, e.g., $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}$. Then $\mathbf{A}\mathbf{B}^+\mathbf{A} \overset{\text{L}}{\leq} \mathbf{A}$ and $\mathbf{B}\mathbf{A}^+\mathbf{B} \overset{\text{L}}{\leq} \mathbf{B}$ but $\mathbf{A} \neq \mathbf{B}$.

Nevertheless, the following holds.

COROLLARY 3. *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{C}_m^H$ *be such that* $\mathcal{R}(\mathbf{A}) = \mathcal{R}(\mathbf{B})$. *Then* $\mathbf{A} \overset{\circ}{\leq} \mathbf{B}$ *if and only if* $\mathbf{B}^+ \overset{\text{L}}{\leq} \mathbf{A}^+$.

Proof. When $\mathbf{A} \overset{\circ}{\leq} \mathbf{B}$, then $\mathbf{A}\mathbf{B}^+\mathbf{A} \overset{\text{L}}{\leq} \mathbf{A}$, or, equivalently, $\mathbf{A} - \mathbf{A}\mathbf{B}^+\mathbf{A} \in \mathbb{C}_m^{\geq}$. Multiplying $\mathbf{A} - \mathbf{A}\mathbf{B}^+\mathbf{A}$ from both sides with $\mathbf{A}^+$ shows $\mathbf{A}^+ - \mathbf{A}^+\mathbf{A}\mathbf{B}^+\mathbf{A}\mathbf{A}^+ \in \mathbb{C}_m^{\geq}$. But since $\mathcal{R}(\mathbf{A}) = \mathcal{R}(\mathbf{B})$ and $\mathcal{R}(\mathbf{B}) = \mathcal{R}(\mathbf{B}^+)$ we have $\mathbf{A}^+\mathbf{A}\mathbf{B}^+ = \mathbf{B}^+ = \mathbf{B}^+\mathbf{A}\mathbf{A}^+$. Hence we obtain $\mathbf{A}^+ - \mathbf{B}^+ \in \mathbb{C}_m^{\geq}$, i.e., $\mathbf{B}^+ \overset{\text{L}}{\leq} \mathbf{A}^+$. On the other hand, multiplying $\mathbf{A}^+ - \mathbf{B}^+ \in \mathbb{C}_m^{\geq}$ from both sides with $\mathbf{A}$ immediately gives $\mathbf{A}\mathbf{B}^+\mathbf{A} \overset{\text{L}}{\leq} \mathbf{A}$. ∎

The following result on antitonicity of the Moore–Penrose inversion with respect to $\overset{\circ}{\leq}$ is strongly related to Corollary 3 and the results in Baksalary, Nordström, and Styan (1990).

COROLLARY 4. *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{C}_m^H$. *Then* $\mathbf{A} \overset{\circ}{\leq} \mathbf{B}$ *and* $\mathbf{B}^+ \overset{\circ}{\leq} \mathbf{A}^+$ *if and only if* $\mathbf{A}^+ \overset{\text{L}}{\leq} \mathbf{B}^+$ *and* $\mathbf{B}^+ \overset{\text{L}}{\leq} \mathbf{A}^+$.

Proof. Suppose $\mathbf{A} \overset{\circ}{\leq} \mathbf{B}$ and $\mathbf{B}^+ \overset{\circ}{\leq} \mathbf{A}^+$. Then $\mathcal{R}(\mathbf{A}) = \mathcal{R}(\mathbf{B})$ and from Corollary 3 $\mathbf{B}^+ \overset{\text{L}}{\leq} \mathbf{A}^+$ and $\mathbf{A}^+ \overset{\text{L}}{\leq} \mathbf{B}^+$.

Suppose, on the other hand, $\mathbf{A}^+ \overset{\text{L}}{\leq} \mathbf{B}^+$ and $\mathbf{B}^+ \overset{\text{L}}{\leq} \mathbf{A}^+$. Then $\mathcal{R}(\mathbf{A}) = \mathcal{R}(\mathbf{B})$ from Theorem 2 and Lemma 4 in Baksalary, Nordström, and Styan (1990), and Corollary 3 gives $\mathbf{B} \overset{\circ}{\leq} \mathbf{A}$ and $\mathbf{A} \overset{\circ}{\leq} \mathbf{B}$. ∎

Recall that two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{C}_{m,n}$ are called parallel summable if the product $\mathbf{A}(\mathbf{A}+\mathbf{B})^-\mathbf{B}$ does not depend on the choice of $(\mathbf{A}+\mathbf{B})^-$ in which case $P(\mathbf{A}, \mathbf{B}) = \mathbf{A}(\mathbf{A}+\mathbf{B})^-\mathbf{B}$ is called the parallel sum of $\mathbf{A}$ and $\mathbf{B}$; cf. section 10.1.6 in Rao and Mitra (1971).

COROLLARY 5. *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{C}_m^H$. *Then* $\mathbf{A}$ *and* $\mathbf{B}$ *are parallel summable and* $P(\mathbf{A}, \mathbf{B}) \in \mathbb{C}_m^{\geq}$ *if and only if* $\mathbf{A} \overset{\circ}{\leq} \mathbf{A} + \mathbf{B}$.

Proof. In view of statement (10.1.29) in Rao and Mitra (1971), $\mathbf{A}$ and $\mathbf{B}$ are parallel summable if and only if $\mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{A} + \mathbf{B})$. Moreover, $P(\mathbf{A}, \mathbf{B}) = \mathbf{A}(\mathbf{A} + \mathbf{B})^+\mathbf{B} = \mathbf{A}(\mathbf{A} + \mathbf{B})^+(\mathbf{A} + \mathbf{B} - \mathbf{A}) = \mathbf{A} - \mathbf{A}(\mathbf{A} + \mathbf{B})^+\mathbf{A} \in \mathbb{C}_m^{\geq}$ if and only if $\mathbf{A}(\mathbf{A} + \mathbf{B})^+\mathbf{A} \overset{\mathrm{L}}{\leq} \mathbf{A}$. □

3. Shorted matrix. Consider the class $\mathcal{A} = \{\mathbf{A} : \mathbf{A} \overset{\circ}{\leq} \mathbf{B}, \mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{C})\}$, where $\mathbf{A}, \mathbf{B} \in \mathbb{C}_m^H$ and $\mathbf{C} \in \mathbb{C}_{m,n}$.

Here we are interested in finding the maximal element in $\mathcal{A}$, which may be called the shorted matrix of $\mathbf{B}$ with respect to $\mathbf{C}$ in $\mathcal{A}$ (in the ordering $\overset{\circ}{\leq}$). Anderson (1971) investigated the class $\mathcal{A}$ as a subclass of $\mathbb{C}_m^{\geq}$ with $\mathbf{B} \in \mathbb{C}_m^{\geq}$, implying that $\overset{\circ}{\leq}$ coincides with $\overset{\mathrm{L}}{\leq}$. The author has shown that $\mathcal{A}$ has a maximal element which is unique.

In the following we demonstrate that this result remains true when $\mathbf{A}$ and $\mathbf{B}$ are considered to be Hermitian. However, we have to impose some restrictions on the subspace $\mathcal{R}(\mathbf{C})$, which are known to be automatically satisfied when $\mathbf{B}$ is a nonnegative definite matrix.

For a matrix $\mathbf{C} \in \mathbb{C}_{m,n}$ we will denote the orthogonal projector onto $\mathcal{R}^{\perp}(\mathbf{C})$ by $\mathbf{Q}_{\mathbf{C}}$, i.e., $\mathbf{Q}_{\mathbf{C}} = \mathbf{I}_m - \mathbf{C}\mathbf{C}^+$. The following lemma might also be of some interest on its own.

LEMMA 1. *Let* $\mathbf{B} \in \mathbb{C}_m^H$ *and* $\mathbf{C} \in \mathbb{C}_{m,n}$. *Then* $\mathcal{R}(\mathbf{B}) + \mathcal{R}(\mathbf{C}) = \mathcal{R}(\mathbf{B}\mathbf{Q}_{\mathbf{C}}) \oplus \mathcal{R}(\mathbf{C})$ *if and only if* $\mathrm{rk}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}) = \mathrm{rk}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})$.

Proof. Clearly $\mathcal{R}(\mathbf{C}) + \mathcal{R}(\mathbf{B}\mathbf{Q}_{\mathbf{C}}) \subseteq \mathcal{R}(\mathbf{C}) + \mathcal{R}(\mathbf{B})$, which means that $\mathcal{R}(\mathbf{C}) + \mathcal{R}(\mathbf{B}\mathbf{Q}_{\mathbf{C}}) = \mathcal{R}(\mathbf{C}) + \mathcal{R}(\mathbf{B})$ if and only if $\mathrm{rk}(\mathbf{C} : \mathbf{B}\mathbf{Q}_{\mathbf{C}}) = \mathrm{rk}(\mathbf{C} : \mathbf{B})$. In view of Theorem 19 in Marsaglia and Styan (1974) the latter is equivalent to $\mathrm{rk}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}) = \mathrm{rk}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})$.

Moreover, $\dim[\mathcal{R}(\mathbf{C}) \cap \mathcal{R}(\mathbf{B}\mathbf{Q}_{\mathbf{C}})] = \mathrm{rk}(\mathbf{C}) + \mathrm{rk}(\mathbf{B}\mathbf{Q}_{\mathbf{C}}) - \mathrm{rk}(\mathbf{C} : \mathbf{B}\mathbf{Q}_{\mathbf{C}})$. By applying again Theorem 19 in Marsaglia and Styan (1974) we obtain $\mathrm{rk}(\mathbf{C} : \mathbf{B}\mathbf{Q}_{\mathbf{C}}) = \mathrm{rk}(\mathbf{C}) + \mathrm{rk}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})$. But when $\mathrm{rk}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}}) = \mathrm{rk}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}) = \mathrm{rk}(\mathbf{B}\mathbf{Q}_{\mathbf{C}})$ we get $\dim[\mathcal{R}(\mathbf{C}) \cap \mathcal{R}(\mathbf{B}\mathbf{Q}_{\mathbf{C}})] = 0$ and hence $\mathcal{R}(\mathbf{C}) + \mathcal{R}(\mathbf{B}\mathbf{Q}_{\mathbf{C}}) = \mathcal{R}(\mathbf{C}) \oplus \mathcal{R}(\mathbf{B}\mathbf{Q}_{\mathbf{C}})$. □

Observe that the condition $\mathrm{rk}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}) = \mathrm{rk}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})$ is equivalent to $\mathcal{R}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}) = \mathcal{R}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})$, since always $\mathcal{R}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}}) \subseteq \mathcal{R}(\mathbf{Q}_{\mathbf{C}}\mathbf{B})$.

When $\mathbf{B} \in \mathbb{C}_m^{\geq}$ then $\mathrm{rk}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}) = \mathrm{rk}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})$ is obviously satisfied and Lemma 1 gives a well-known result; cf. Lemma 2.1 in Rao (1974).

COROLLARY 6. *Let* $\mathbf{B} \in \mathbb{C}_m^H$ *and* $\mathbf{C} \in \mathbb{C}_{m,n}$ *be such that* $\mathcal{R}(\mathbf{B}) + \mathcal{R}(\mathbf{C}) = \mathcal{R}(\mathbf{B}\mathbf{Q}_{\mathbf{C}}) \oplus \mathcal{R}(\mathbf{C})$. *Then* $\dim[\mathcal{R}(\mathbf{B}) \cap \mathcal{R}(\mathbf{C})] = \mathrm{rk}(\mathbf{B}) - \mathrm{rk}(\mathbf{B}\mathbf{Q}_{\mathbf{C}})$.

Proof. Observe that $\dim[\mathcal{R}(\mathbf{C}) \cap \mathcal{R}(\mathbf{B})] = \mathrm{rk}(\mathbf{C}) + \mathrm{rk}(\mathbf{B}) - \mathrm{rk}(\mathbf{C} : \mathbf{B})$. In view of Theorem 19 in Marsaglia and Styan (1974) we have $\mathrm{rk}(\mathbf{C} : \mathbf{B}) = \mathrm{rk}(\mathbf{C}) + \mathrm{rk}(\mathbf{Q}_{\mathbf{C}}\mathbf{B})$. This immediately leads to the assertion. □

LEMMA 2. *Let* $\mathbf{B} \in \mathbb{C}_m^H$ *and* $\mathbf{C} \in \mathbb{C}_{m,n}$ *be such that* $\mathcal{R}(\mathbf{B}) + \mathcal{R}(\mathbf{C}) = \mathcal{R}(\mathbf{B}\mathbf{Q}_{\mathbf{C}}) \oplus \mathcal{R}(\mathbf{C})$. *Let* $\mathbf{S} = \mathbf{B} - \mathbf{B}\mathbf{Q}_{\mathbf{C}}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})^+\mathbf{Q}_{\mathbf{C}}\mathbf{B}$. *Then the following two statements hold*:

(a) $\mathbf{S} \overset{\mathrm{rs}}{\leq} \mathbf{B}$, *i.e.;* $\mathcal{R}(\mathbf{S}) \subseteq \mathcal{R}(\mathbf{B})$ *and* $\mathbf{S}\mathbf{B}^+\mathbf{S} = \mathbf{S}$;

(b) $\mathrm{rk}(\mathbf{S}) = \dim[\mathcal{R}(\mathbf{B}) \cap \mathcal{R}(\mathbf{C})]$, *or, equivalently,* $\mathcal{R}(\mathbf{S}) = \mathcal{R}(\mathbf{B}) \cap \mathcal{R}(\mathbf{C})$.

Proof. In view of the assumptions we have $\mathcal{R}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}) = \mathcal{R}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})$, which immediately gives

$$\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})^+\mathbf{Q}_{\mathbf{C}}\mathbf{B} = \mathbf{Q}_{\mathbf{C}}\mathbf{B}. \tag{5}$$

To observe statement (a) we only have to show $\mathbf{S}\mathbf{B}^+\mathbf{S} = \mathbf{S}$ since $\mathcal{R}(\mathbf{S}) \subseteq \mathcal{R}(\mathbf{B})$

is obviously satisfied. Now, since $\mathbf{B}\mathbf{B}^{+}\mathbf{S} = \mathbf{S}$ we immediately get $\mathbf{S}\mathbf{B}^{+}\mathbf{S} = \mathbf{S} - \mathbf{B}\mathbf{Q}_{\mathbf{C}}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})^{+}\mathbf{Q}_{\mathbf{C}}\mathbf{S}$. Since in view of (5) $\mathbf{Q}_{\mathbf{C}}\mathbf{S} = \mathbf{0}$ we obtain $\mathbf{S}\mathbf{B}^{+}\mathbf{S} = \mathbf{S}$.

Moreover, $\mathrm{rk}(\mathbf{S}) = \mathrm{rk}[\mathbf{B}(\mathbf{I}_m - \mathbf{P})]$, where $\mathbf{P} := \mathbf{Q}_{\mathbf{C}}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})^{+}\mathbf{Q}_{\mathbf{C}}\mathbf{B}$ is idempotent, or, in other words, $\mathbf{P}$ is a projector. Applying Corollary 6.2 in Marsaglia and Styan (1974) shows that

$$\mathrm{rk}(\mathbf{S}) = \mathrm{rk}[\mathbf{B}(\mathbf{I}_m - \mathbf{P})] = \mathrm{rk}(\mathbf{I}_m - \mathbf{P}) - \dim[\mathcal{N}(\mathbf{B}) \cap \mathcal{R}(\mathbf{I}_m - \mathbf{P})].$$

But since $\mathbf{P}$ is a projector we have $\mathcal{R}(\mathbf{I}_m - \mathbf{P}) = \mathcal{N}(\mathbf{P})$; see, e.g., Theorem 3.6.3 in Rao and Bhimasankaram (1992). By using (5) we immediately get $\mathcal{N}(\mathbf{P}) = \mathcal{N}(\mathbf{Q}_{\mathbf{C}}\mathbf{B})$, and hence $\mathcal{N}(\mathbf{B}) \cap \mathcal{R}(\mathbf{I}_m - \mathbf{P}) = \mathcal{N}(\mathbf{B})$. This gives $\dim[\mathcal{N}(\mathbf{B}) \cap \mathcal{R}(\mathbf{I}_m - \mathbf{P})] = m - \mathrm{rk}(\mathbf{B})$. Since $\mathbf{P}$ is a projector we also have $\mathrm{rk}(\mathbf{I}_m - \mathbf{P}) = m - \mathrm{rk}(\mathbf{P})$. See again Theorem 3.6.3 in Rao and Bhimasankaram (1992), which leads to

$$\mathrm{rk}(\mathbf{S}) = \mathrm{rk}(\mathbf{B}) - \mathrm{rk}(\mathbf{P}).$$

In view of (5) we have $\mathrm{rk}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}) = \mathrm{rk}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})^{+}\mathbf{Q}_{\mathbf{C}}\mathbf{B}) = \mathrm{rk}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{P}) \le \mathrm{rk}(\mathbf{P}) \le \mathrm{rk}(\mathbf{Q}_{\mathbf{C}}\mathbf{B})$, which gives $\mathrm{rk}(\mathbf{P}) = \mathrm{rk}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}) = \mathrm{rk}(\mathbf{B}\mathbf{Q}_{\mathbf{C}})$. Hence, $\mathrm{rk}(\mathbf{S}) = \dim[\mathcal{R}(\mathbf{B}) \cap \mathcal{R}(\mathbf{C})]$ follows from Corollary 6.

By noting that $\mathbf{B}\mathbf{B}^{+}\mathbf{S} = \mathbf{S}$ and $\mathbf{S}\mathbf{Q}_{\mathbf{C}} = \mathbf{0}$, i.e., $\mathcal{R}(\mathbf{S}) \subseteq \mathcal{R}(\mathbf{B}) \cap \mathcal{R}(\mathbf{C})$, we obviously get the equivalence of $\mathrm{rk}(\mathbf{S}) = \dim[\mathcal{R}(\mathbf{B}) \cap \mathcal{R}(\mathbf{C})]$ and $\mathcal{R}(\mathbf{S}) = \mathcal{R}(\mathbf{B}) \cap \mathcal{R}(\mathbf{C})$, and statement (b) is shown. □

Observe that under the assumption $\mathcal{R}(\mathbf{B}) + \mathcal{R}(\mathbf{C}) = \mathcal{R}(\mathbf{B}\mathbf{Q}_{\mathbf{C}}) \oplus \mathcal{R}(\mathbf{C})$ the matrix $\mathbf{S}$ in Lemma 2 does not depend on the choice of generalized inverse $(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})^{-}$, i.e., $\mathbf{S} = \mathbf{B} - \mathbf{B}\mathbf{Q}_{\mathbf{C}}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})^{-}\mathbf{Q}_{\mathbf{C}}\mathbf{B}$ for any $(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})^{-}$. However, we may also write $\mathbf{S} = \mathbf{B} - \mathbf{B}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})^{+}\mathbf{B}$, since $(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})^{+} = (\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})^{+}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})^{+}\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}}$, i.e., $(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})^{+} = (\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})^{+}\mathbf{Q}_{\mathbf{C}} = \mathbf{Q}_{\mathbf{C}}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})^{+}$. It should be pointed out that the plus superscript in

$$\mathbf{S} = \mathbf{B} - \mathbf{B}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})^{+}\mathbf{B}$$

cannot be replaced by a minus superscript unless $\mathcal{R}(\mathbf{B}) \subseteq \mathcal{R}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})$. The latter inclusion is equivalent to $\mathcal{R}(\mathbf{B}) \cap \mathcal{R}(\mathbf{C}) = \{\mathbf{0}\}$ when $\mathcal{R}(\mathbf{B}) + \mathcal{R}(\mathbf{C}) = \mathcal{R}(\mathbf{B}\mathbf{Q}_{\mathbf{C}}) \oplus \mathcal{R}(\mathbf{C})$ is satisfied. In view of Lemma 2, $\mathcal{R}(\mathbf{B}) \cap \mathcal{R}(\mathbf{C}) = \{\mathbf{0}\}$ means $\mathbf{S} = \mathbf{0}$.

We are now ready to present the main result of this section. It is not only concerned with the above introduced class $\mathcal{A} = \{\mathbf{A} \in \mathbb{C}_m^H : \mathbf{A} \overset{\circ}{\le} \mathbf{B}, \mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{C})\}$ but also with the class $\mathcal{A}_{\mathrm{rs}} = \{\mathbf{A} \in \mathbb{C}_m^H : \mathbf{A} \overset{\mathrm{rs}}{\le} \mathbf{B}, \mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{C})\}$.

THEOREM 2. *Let $\mathbf{B} \in \mathbb{C}_m^H$ and $\mathbf{C} \in \mathbb{C}_{m,n}$ be such that $\mathcal{R}(\mathbf{B}) + \mathcal{R}(\mathbf{C}) = \mathcal{R}(\mathbf{B}\mathbf{Q}_{\mathbf{C}}) \oplus \mathcal{R}(\mathbf{C})$. Then $\mathbf{S} = \mathbf{B} - \mathbf{B}\mathbf{Q}_{\mathbf{C}}(\mathbf{Q}_{\mathbf{C}}\mathbf{B}\mathbf{Q}_{\mathbf{C}})^{+}\mathbf{Q}_{\mathbf{C}}\mathbf{B}$ is the maximal element in $\mathcal{A}$ as well as in $\mathcal{A}_{\mathrm{rs}}$.*

Proof. In view of Lemma 2 the matrix $\mathbf{S}$ belongs to $\mathcal{A}_{\mathrm{rs}}$ and hence also to $\mathcal{A}$.

To demonstrate that $\mathbf{S}$ is maximal in $\mathcal{A}$ consider an arbitrary matrix $\mathbf{A} \in \mathbb{C}_m^H$ satisfying $\mathbf{A} \overset{\circ}{\le} \mathbf{B}$ and $\mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{C})$. Then obviously $\mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{B}) \cap \mathcal{R}(\mathbf{C})$, where $\mathcal{R}(\mathbf{B}) \cap \mathcal{R}(\mathbf{C}) = \mathcal{R}(\mathbf{S})$ in view of Lemma 2. Since $\mathbf{S} \overset{\mathrm{rs}}{\le} \mathbf{B}$, we have $\mathbf{S}\mathbf{B}^{+}\mathbf{S} = \mathbf{S}$. Multiplying from the left with $\mathbf{A}\mathbf{S}^{+}$ and from the right with $(\mathbf{A}\mathbf{S}^{+})^{*} = \mathbf{S}^{+}\mathbf{A}$ gives $\mathbf{A}\mathbf{S}^{+}\mathbf{S}\mathbf{B}^{+}\mathbf{S}\mathbf{S}^{+}\mathbf{A} = \mathbf{A}\mathbf{S}^{+}\mathbf{S}\mathbf{S}^{+}\mathbf{A} = \mathbf{A}\mathbf{S}^{+}\mathbf{A}$. In view of $\mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{S})$ we have $\mathbf{A}\mathbf{S}^{+}\mathbf{S} = \mathbf{A} = \mathbf{S}\mathbf{S}^{+}\mathbf{A}$, showing that $\mathbf{A}\mathbf{B}^{+}\mathbf{A} = \mathbf{A}\mathbf{S}^{+}\mathbf{A}$. Hence, $\mathbf{A} \overset{\circ}{\le} \mathbf{B}$ implies $\mathbf{A} \overset{\circ}{\le} \mathbf{S}$, showing that $\mathbf{S}$ is maximal in $\mathcal{A}$.

When $\mathbf{A}$ is an arbitrary matrix in $\mathcal{A}_{\mathrm{rs}}$, then we also obtain $\mathbf{A}\mathbf{B}^{+}\mathbf{A} = \mathbf{A}\mathbf{S}^{+}\mathbf{A}$. Hence, $\mathbf{A} \overset{\mathrm{rs}}{\le} \mathbf{B}$ implies $\mathbf{A} \overset{\mathrm{rs}}{\le} \mathbf{S}$, showing that $\mathbf{S}$ is maximal in $\mathcal{A}_{\mathrm{rs}}$. □

As mentioned above, our Theorem 2 generalizes results of Anderson (1971, Theorem 1), Goller (1986, Theorem 4.1), and Mitra, Puntanen, and Styan (1994, Theorem 3.3), who considered maximal elements in subclasses of $\mathbb{C}_m^{\geq}$. Clearly we automatically have $\mathcal{R}(\mathbf{B}) + \mathcal{R}(\mathbf{C}) = \mathcal{R}(\mathbf{B}\mathbf{Q}_{\mathbf{C}}) \oplus \mathcal{R}(\mathbf{C})$ and $\mathbf{A} \overset{\circ}{\leq} \mathbf{B} \Leftrightarrow \mathbf{A} \overset{\mathrm{L}}{\leq} \mathbf{B}$ when $\mathbf{A}, \mathbf{B} \in \mathbb{C}_m^{\geq}$.

Note that, e.g., Mitra and Puri (1982), Mitra (1986), and Goller (1986) also consider maximal elements in subclasses of $\mathbb{C}_{m,n}$.

4. Concluding remarks. Observe that any partial ordering in the set of complex Hermitian matrices can be extended to the broader set of complex square matrices. For this let $H(\mathbf{A})$ denote the Hermitian part of $\mathbf{A} \in \mathbb{C}_m$, i.e., $H(\mathbf{A}) = \frac{1}{2}(\mathbf{A}+\mathbf{A}^*)$, and $S(\mathbf{A})$ denote the skew-Hermitian part of $\mathbf{A}$, i.e., $S(\mathbf{A}) = \frac{1}{2}(\mathbf{A} - \mathbf{A}^*)$. Then $\mathbf{A} = H(\mathbf{A}) + S(\mathbf{A})$.

Now, when $\overset{?}{\leq}$ denotes an arbitrary partial ordering defined for Hermitian matrices only, specify the extended partial ordering in $\mathbb{C}_m$ by

$$\mathbf{A} \overset{?}{\leq} \mathbf{B} \quad :\Leftrightarrow \quad H(\mathbf{A}) \overset{?}{\leq} H(\mathbf{B}) \text{ and } S(\mathbf{A}) = S(\mathbf{B}).$$

It is clear that $S(\mathbf{A}) = S(\mathbf{B})$ is satisfied if and only if the difference $\mathbf{B}-\mathbf{A}$ is Hermitian. Note that the well-known Löwner partial ordering defined by (2) may be characterized in this way; i.e., we have

$$\mathbf{A} \overset{\mathrm{L}}{\leq} \mathbf{B} \quad \Leftrightarrow \quad H(\mathbf{A}) \overset{\mathrm{L}}{\leq} H(\mathbf{B}) \text{ and } S(\mathbf{A}) = S(\mathbf{B}).$$

REFERENCES

[1] W. N. Anderson Jr. (1971), *Shorted operators*, SIAM J. Appl. Math., 20, pp. 520–525.

[2] J. K. Baksalary, K. Nordström, and G. P. H. Styan (1990), *Löwner-ordering antitonicity of generalized inverses of Hermitian matrices*, Linear Algebra Appl., 127, pp. 171–182.

[3] J. K. Baksalary, F. Pukelsheim, and G. P. H. Styan (1989), *Some properties of matrix partial orderings*, Linear Algebra Appl., 119, pp. 57–85.

[4] H. Goller (1986), *Shorted operators and rank decomposition matrices*, Linear Algebra Appl., 81, pp. 207–236.

[5] J. Gross (1997), *Some remarks on partial orderings of Hermitian matrices*, Linear and Multilinear Algebra, 42, pp. 53–60.

[6] G. Marsaglia and G. P. H. Styan (1974), *Equalities and inequalities for ranks of matrices*, Linear and Multilinear Algebra, 2, pp. 269–292.

[7] S. K. Mitra (1986), *The minus partial order and the shorted matrix*, Linear Algebra Appl., 83, pp. 1–27.

[8] S. K. Mitra, S. Puntanen and G. P. H. Styan (1994), *Shorted Matrices and Their Applications in Linear Statistical Models: A Review*, Report A 287, Department of Mathematical Sciences, University of Tampere, Tampere, Finland.

[9] S. K. Mitra and M. L. Puri (1982), *Shorted matrices—An extended concept and some applications*, Linear Algebra Appl., 42, pp. 57–79.

[10] A. R. Rao and P. Bhimasankaram (1992), *Linear Algebra*, Tata McGraw–Hill, New Delhi, India.

[11] C. R. Rao (1974), *Projectors, generalized inverses and the BLUE's*, J. Roy. Statist. Soc. Ser. B, 36, pp. 442–448.

[12] C. R. Rao and S. K. Mitra (1971), *Generalized Inverse of Matrices and its Applications*, Wiley, New York.

[13] G. P. H. Styan (1985), *Schur complements and linear statistical models*, in Proc. 1st Internat. Tampere Seminar Linear Models Appl., T. Pukkila and S. Puntanen, eds., Department of Mathematical Sciences, University of Tampere, Tampere, Finland, pp. 37–75.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 4, pp. 893–912, October 1997

008

A UNIFIED REPRESENTATION AND THEORY OF ALGEBRAIC ADDITIVE SCHWARZ AND MULTISPLITTING METHODS*

ANDREAS FROMMER† AND HARTMUT SCHWANDT‡

Abstract. We develop a unified representation of two well-known approaches for the solution of linear systems of equations by partitioning the original system into overlapping subsystems. The representation generalizes the algebraic form of both the additive Schwarz and multisplitting methods. In the new formulation we obtain convergence results similar to those known for multisplittings, considering one- and two-stage variants. We report on some numerical experiments on a CRAY T3D which suggest a slight preference for algebraic additive Schwarz methods over multisplitting methods. These experiments also demonstrate the efficiency of our approach in a parallel computing environment.

Key words. multisplittings, additive Schwarz methods, nonnegative matrices

AMS subject classifications. 65F10, 65Y05

PII. S0895479896301212

1. Introduction.

The classical Schwarz alternating procedure (SAP) has been introduced in [29] in the last century in order to provide a constructive proof for the existence of a (continuous) solution of a class of partial elliptic boundary value problems. The domain of integration is partitioned into possibly overlapping subdomains. Starting from an initial approximation to the solution, one iteratively solves a smaller boundary value problem alternatingly for each subdomain. Due to the domain decomposition, "artificial" boundaries are introduced on each subdomain. These "artificial" boundaries are updated iteratively by appropriate parts of the approximations on the neighboring subdomains. The SAP represents one of the best known and most widely used domain decomposition principles which has been used in many methods for both the continuous and the discrete solution not only of elliptic, but also of parabolic and hyperbolic, partial differential equations; see, e.g., [4], [14], [17], [18], [28], [33]. The original Schwarz method can be classified as an example of a *multiplicative* Schwarz method. In the so-called *additive* Schwarz methods the alternating treatment of the subproblems in each iterative step is replaced by a simultaneous treatment. Additive Schwarz-type methods have proven to be very useful in parallel contexts, particularly as a preconditioning technique (see [4] and the references therein).

The idea of the additive Schwarz method has been applied to general algebraic systems of linear equations by several authors; see [6], [25], [26], [27]. These *algebraic* additive Schwarz methods are related—but not identical—to those from domain decomposition. The idea of overlapping subdomains is now reflected by overlapping subsystems. This provides the connection to multisplitting methods which result from an entirely algebraic background. In the case of a linear system $Ax = b$ with a nonsingular coefficient matrix A, multisplittings cover certain block splittings of A with overlapping diagonal blocks. Analogously to the additive Schwarz principle, the overlaps are introduced to optimize the relation of asymptotic convergence speed

* Received by the editors March 27, 1996; accepted for publication (in revised form) by D. P. O'Leary October 8, 1996.
http://www.siam.org/journals/simax/18-4/30121.html
† Fachbereich Mathematik, Bergische Universität GH Wuppertal, D-42097 Wuppertal, Germany (frommer@math.uni-wuppertal.de).
‡ Fachbereich Mathematik, Technische Universität Berlin, D-10623 Berlin, Germany (schwandt@math.tu-berlin.de).

and arithmetic work per iteration and to improve the load balancing for a parallel solution. Multisplitting methods have been introduced in [21]. Among the numerous papers which have appeared since then, we note, e.g., [1], [5], [8], [9], [10], [15], [20], [30].

In this paper we present a general model which covers both algebraic additive Schwarz methods and multisplitting methods. As opposed to [6], [25], [26], [27], where special sparsity patterns (like block tridiagonality) are required, our model does not rely on any particular structure of the system to be solved. We develop a unified convergence analysis for our general model and, moreover, stress the strong similarity of the resulting algorithmic formulations for multisplitting and algebraic additive Schwarz methods.

The paper is organized as follows: in section 2 we develop and motivate our concept starting from an example. Section 3 contains basic definitions and auxiliary results, whereas in section 4 we prove several convergence results based on the theory of nonnegative matrices and H-matrices. Section 5 considers two-stage ("inner-outer") modifications which are relevant for practical applications and for which we present convergence results similar to those in section 4. In section 6 we give some comparison results with methods based on a single splitting. Section 7 discusses algorithmic aspects. In section 8 we report the results of some numerical experiments.

2. The representation. We develop the idea behind both types of methods by the example of two overlapping index sets which are the algebraic counterpart of two overlapping subdomains in the case of a SAP applied to a continuous boundary value problem. Let

$$Ax = b, \quad A \in \mathbf{R}^{n\times n}, \;\; x, b \in \mathbf{R}^n \;,$$

be a linear system with nonsingular coefficient matrix A and let $S, \widetilde{S}$ be a decomposition of the set $\{1, \ldots, n\}$; i.e., $\emptyset \neq S, \widetilde{S}, \;\; S \cup \widetilde{S} = \{1, \ldots, n\}$, and $O := S \cap \widetilde{S} \neq \emptyset$. For ease of notation we assume that $S, \widetilde{S}$ consist of blocks of consecutive indices, $S = \{1, \ldots, m\}$, $\widetilde{S} = \{\widetilde{m}, \ldots, n\}$ with $\widetilde{m} \leq m$, but this is by no means mandatory. The overlapping part is thus given by $O = \{\widetilde{m}, \ldots, m\}$. We decompose A as a 3×3 block matrix:

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix},$$

with $A_{11} \in \mathbf{R}^{(\widetilde{m}-1)\times(\widetilde{m}-1)}$, $A_{22} \in \mathbf{R}^{(m-\widetilde{m}+1)\times(m-\widetilde{m}+1)}$, $A_{33} \in \mathbf{R}^{(n-m)\times(n-m)}$, and

$$b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}, \quad b_1 \in \mathbf{R}^{\widetilde{m}-1}, \, b_2 \in \mathbf{R}^{m-\widetilde{m}+1}, \, b_3 \in \mathbf{R}^{n-m} \;.$$

Abusing notation, we will also write $b = (b_1, b_2, b_3)^T$ instead of the correct but more complicated $(b_1^T, b_2^T, b_3^T)^T$. Given an initial approximation $x^0 = (x_1^0, x_2^0, x_3^0)^T$, we now calculate first iterates $(x_1^1, x_2^1)^T$ and $(\tilde{x}_2^1, \tilde{x}_3^1)^T$ for both subsets $S, \widetilde{S}$ by solving

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x_1^1 \\ y_2^1 \end{pmatrix} = \begin{pmatrix} b_1 - A_{13}x_3^0 \\ b_2 - A_{23}x_3^0 \end{pmatrix} \;,$$
$$\begin{pmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{pmatrix} \begin{pmatrix} \widetilde{y}_2^1 \\ \widetilde{x}_3^1 \end{pmatrix} = \begin{pmatrix} b_2 - A_{21}x_1^0 \\ b_3 - A_{31}x_1^0 \end{pmatrix} \;,$$

yielding two potentially *different* approximations $y_2^1, \widetilde{y}_2^1$ on the overlapping part O. Taking nonnegative diagonal weighting matrices $\underline{E}_1, \underline{E}_2, \widetilde{\underline{E}}_1, \widetilde{\underline{E}}_2$ of dimension $(m - \widetilde{m} + 1)$ with $\underline{E}_1 + \underline{E}_2 = \widetilde{\underline{E}}_1 + \widetilde{\underline{E}}_2 = I$ we obtain a variety of possibilities for keeping different and/or weighted approximations on the overlap by setting

$$x_2^1 = \underline{E}_1 y_2^1 + \underline{E}_2 \widetilde{y}_2^1,$$
$$\widetilde{x}_2^1 = \widetilde{\underline{E}}_1 \widetilde{y}_2^1 + \widetilde{\underline{E}}_2 y_2^1.$$

In general, the $(k+1)$st iteration step is given by

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x_1^{k+1} \\ y_2^{k+1} \end{pmatrix} = \begin{pmatrix} b_1 - A_{13}\widetilde{x}_3^k \\ b_2 - A_{23}\widetilde{x}_3^k \end{pmatrix}, \tag{1}$$

$$\begin{pmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{pmatrix} \begin{pmatrix} \widetilde{y}_2^{k+1} \\ \widetilde{x}_3^{k+1} \end{pmatrix} = \begin{pmatrix} b_2 - A_{21}x_1^k \\ b_3 - A_{31}x_1^k \end{pmatrix}, \tag{2}$$

and

$$x_2^{k+1} = \underline{E}_1 y_2^{k+1} + \underline{E}_2 \widetilde{y}_2^{k+1}, \tag{3}$$

$$\widetilde{x}_2^{k+1} = \widetilde{\underline{E}}_1 \widetilde{y}_2^{k+1} + \widetilde{\underline{E}}_2 y_2^{k+1}. \tag{4}$$

We thus obtain an approximation x_1^{k+1} for the solution on $S_1 \backslash O$, $\widetilde{x}_3^{k+1}$ on $S_2 \backslash O$, and two possibly different approximations $x_2^{k+1}, \widetilde{x}_2^{k+1}$ on the overlapping part O. In the algebraic additive Schwarz method one takes

$$\underline{E}_1 = I,\ \underline{E}_2 = 0,\ \widetilde{\underline{E}}_1 = 0,\ \widetilde{\underline{E}}_2 = I,\ \text{i.e.,}\ \ x_2^{k+1} = y_2^{k+1},\ \widetilde{x}_2^{k+1} = \widetilde{y}_2^{k+1},$$

while multisplitting methods are characterized by

$$\underline{E}_1 = \widetilde{\underline{E}}_1,\ \ \underline{E}_2 = \widetilde{\underline{E}}_2,\ \ \text{i.e.,}\ \ \widetilde{x}_2^{k+1} = x_2^{k+1} = \underline{E}_1 y_2^{k+1} + \underline{E}_2 \widetilde{y}_2^{k+1}.$$

If we have more than two overlapping subsets which are not necessarily blocks of consecutive indices, it becomes increasingly cumbersome to formulate the method in a way similar to (1)–(4). This is the reason we now introduce blocks x_3^k and $\widetilde{x}_1^k$ which can be treated as *dummy* blocks, which will never have to be computed in practice in the above context. We set

$$M = \begin{pmatrix} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} & 0 \\ 0 & 0 & A_{33} \end{pmatrix}, \qquad N = M - A,$$

$$\widetilde{M} = \begin{pmatrix} A_{11} & 0 & 0 \\ 0 & A_{22} & A_{23} \\ 0 & A_{32} & A_{33} \end{pmatrix}, \qquad \widetilde{N} = \widetilde{M} - A,$$

and $x^k = (x_1^k, x_2^k, x_3^k)$, $\widetilde{x}^k = (\widetilde{x}_1^k, \widetilde{x}_2^k, \widetilde{x}_3^k)$. By defining

$$E_1 = \begin{pmatrix} I & & \\ & \underline{E}_1 & \\ & & 0 \end{pmatrix}, \qquad E_2 = \begin{pmatrix} I & & \\ & \underline{E}_2 & \\ & & 0 \end{pmatrix},$$

and similarly, $\widetilde{E}_1, \widetilde{E}_2$, we can reformulate (1)–(4) as

$$x^{k+1} = E_1(M^{-1}(Nx^k + b)) + E_2(\widetilde{M}^{-1}(\widetilde{N}\widetilde{x}^k + b)), \tag{5}$$

$$\widetilde{x}^{k+1} = \widetilde{E}_1(M^{-1}(Nx^k + b)) + \widetilde{E}_2(\widetilde{M}^{-1}(\widetilde{N}\widetilde{x}^k + b)). \tag{6}$$

Due to the block structure of $M, N, \widetilde{M}, \widetilde{N}$ and the weighting matrices $E_1, E_2, \widetilde{E}_1, \widetilde{E}_2$, neither x_3^{k+1} nor $\widetilde{x}_1^{k+1}$ in (5), (6) contribute to x_1^{k+2}, x_2^{k+2} or $\widetilde{x}_2^{k+2}, \widetilde{x}_3^{k+2}$ in the next iterative step. Also note that we could replace A_{33} in M and A_{11} in $\widetilde{M}$ by any other (nonsingular) matrix without any influence on the computed blocks $x_1^{k+1}, x_2^{k+1}, \widetilde{x}_2^{k+1}, \widetilde{x}_3^{k+1}$.

Besides its notational simplicity, the major advantage of the formulation of (5), (6) lies in the fact that we now deal with "ordinary" splittings $A = M - N$, $A = \widetilde{M} - \widetilde{N}$ of the full matrix A. This will make theoretical investigations much easier. Generalizing (5) and (6) to more than two splittings yields the following fundamental definition.

DEFINITION 2.1. *Let $A \in \mathbf{R}^{n\times n}$ be nonsingular. A collection of L splittings $A = M_l - N_l \in \mathbf{R}^{n\times n}$, $l = 1, \ldots, L$, and L^2 nonnegative diagonal matrices $E_{l,m} \in \mathbf{R}^{n\times n}$ such that $\sum_{m=1}^{L} E_{l,m} = I$ for $l = 1, \ldots, L$ is called a* weighted additive Schwarz-type splitting *of A.*

The corresponding weighted additive Schwarz-type method *computes iterates $x^{k,l}, l = 1, \ldots, L$, by*

$$x^{k+1,l} = \sum_{m=1}^{L} E_{l,m} y^{k,m}, \qquad k = 0, 1, \ldots, \tag{7}$$

where

$$M_m y^{k,m} = N_m x^{k,m} + b, \qquad m = 1, \ldots, L, \tag{8}$$

with a priori given initial approximations $x^{0,l}$, $l = 1, \ldots, L$.

We emphasize that Definition 2.1 does not assume that the matrices M_l are built from blocks of the matrix A nor does it assume any block structure for the M_l at all. However, such block structure, together with the zero pattern of the $E_{l,m}$, will be crucial for the determination of those (block) components of the iterates which will actually have to be computed to perform (7).

Multisplittings can be considered as a special case of weighted additive Schwarz-type methods where $E_{l,m} = E_m$ does not depend on l, so there is a unique iterate which does not depend on l. This is specified in our next definition.

DEFINITION 2.2. *A* multisplitting *of $A \in \mathbf{R}^{n\times n}$ is a collection of L splittings $A = M_l - N_l, l = 1, \ldots, L$, and L nonnegative diagonal matrices E_l such that $\sum_{m=1}^{L} E_m = I$. The corresponding* multisplitting iteration *is defined by*

$$x^{k+1} = \sum_{m=1}^{L} E_m y^{k,m}, \tag{9}$$

where

$$M_m y^{k,m} = N_m x^k + b, \qquad m = 1, \ldots, L\ . \tag{10}$$

We can retrieve the algebraic additive Schwarz methods considered in [6], [25], [26], [27], e.g., as another special case of Definition 2.1. Consider a decomposition of $\{1, \ldots, n\}$ into (overlapping) subsets $S_m, m = 1, \ldots, L$, and define $P_m \in \mathbf{R}^{n\times n}$ to be the projection onto S_m; i.e., P_m is diagonal with its jth diagonal element equal to 1 if $j \in S_m$ and 0 otherwise.

DEFINITION 2.3. *Assume*

$$(E_{l,m})_{ii} \in \{0,1\}, \quad i = 1,\dots,n, \quad l,m = 1,\dots,L, \tag{11}$$

for all diagonal elements of all matrices $E_{l,m}$ *in Definition* 2.1 *and, in addition,*

$$E_{l,l} = P_l, \qquad l = 1,\dots,L. \tag{12}$$

Then, we call the resulting weighted additive Schwarz-type splitting an algebraic additive Schwarz splitting.

By (12) we ensure that for the components of $x^{k+1,l}$ which belong to S_l we just take the result from the lth splitting, regardless of any possible overlap of S_l with other sets S_m. Condition (11) states that the weighting process is specialized to a *choice*.

Different possibilities still exist to define the remaining matrices $E_{l,m}$, subject to the condition $\sum_{m=1,m\neq l}^{L} E_{l,m} = I - P_l$, resulting from (12) and Definition 2.1. For example, we can take

$$E_{l,m} = \begin{cases} (I-P_l)\displaystyle\prod_{i=1}^{m-1}(I-P_i)P_m & \text{for } m < l, \\ \displaystyle\prod_{i=1}^{m-1}(I-P_i)P_m & \text{for } m > l, \end{cases} \tag{13}$$

but various other choices are conceivable. Any such choice may be interpreted as follows: the equation for updating component $i \in S_l$ of $x^{k,l}$ induced by the splitting $M_l - N_l$ will usually involve several components $j \notin S_l$. Such components j can belong to *several* other blocks S_m, and one has to decide which block $S_{m'}$ to take by setting the jth diagonal element of $E_{l,m'}$ to 1. We illustrate this point by the following example. Assume a finite difference discretization of a partial elliptic boundary value problem with a five-point stencil on a rectangular domain which is decomposed vertically and horizontally into rectangular subdomains which overlap on all four edges with the respective neighbor. Then there exist grid points x belonging to four different subdomains, and when evaluating a stencil involving x, one has to decide from which subdomain x is taken.

The general case of Definition 2.1 can be regarded as a hybrid variant of an algebraic additive Schwarz method and a multisplitting method: we do not require the diagonal entries of the $E_{l,m}$ to be equal to either 0 or 1 only. A component belonging to an overlap between several subsets can be chosen as a general convex combination of the individual contributions. On each overlap, different approximations will be kept during the whole iteration and each of these approximations will be itself the result of a weighting process for the approximations on the overlaps from the preceding iteration.

While all three definitions above are very appropriate for theoretical purposes they do not at all address the issue of computational efficiency. In fact, having to compute all components of all vectors $y^{k,m}$ in each iteration would certainly yield a quite inefficient computational process, and our introductory example showed that the vectors $y^{k,m}$ may contain "dummy blocks" which are irrelevant for the iteration and need not be computed.

In order to address this issue in the general setting of Definition 2.1, we define the *block property* of an algebraic additive Schwarz-type splitting as follows.

DEFINITION 2.4. *An algebraic additive Schwarz-type splitting is said to satisfy the* block property *if for all* $l, m = 1, \ldots, L$, *we have*

$$(E_{m,m})_{ii} = 0 \Rightarrow (E_{l,m})_{ii} = 0 \quad \textit{for } i = 1, \ldots, n \ .$$

Any multisplitting, according to Definition 2.2, trivially satisfies the block property. In this case, the component $(y^{k,m})_i$ only needs to be computed if $(E_m)_{ii} \neq 0$. In the general case of Definition 2.1, and so, in particular, for the algebraic additive Schwarz splittings of Definition 2.3, the block property ensures that we only have to compute those components i of $y^{k,m}$ in $M_m y^{k,m} = N_m x^{k,m} + b$ for which $(E_{m,m})_{ii} \neq 0$, since only these components possibly contribute to $x^{k+1,l} = \sum_{m=1}^{L} E_{l,m} y^{k,m}$ for *all* $l = 1, \ldots, L$. For example, the choice for $E_{l,m}$ given in (13) for an algebraic additive Schwarz splitting satisfies the block property.

3. Notation and auxiliary results. In $\mathbf{R}^n$ and $\mathbf{R}^{n\times n}$ the relation $\leq$ denotes the natural componentwise partial ordering. In addition, for $x, y \in \mathbf{R}^n$ we write $x < y$ if $x_i < y_i$, $i = 1, \ldots, n$. A vector $x \geq 0$ is called *nonnegative*; if $x > 0$ we call x *positive*. Similarly, $A \in \mathbf{R}^{n\times n}$ is called *nonnegative* if $A \geq 0$. If A is nonsingular with $A^{-1} \geq 0$ and $u \in \mathbf{R}^n$ is positive, then $A^{-1}u > 0$ since A^{-1} cannot contain a row whose entries are all equal to 0. This fact will be used several times in the sections below without any further comment.

A representation $A = M - N$, $A, M, N \in \mathbf{R}^{n\times n}$ is termed a *splitting* of A if M is nonsingular. A splitting $A = M - N$ is termed *regular* (*weak regular*) if $M^{-1} \geq 0$ and $N \geq 0$ ($M^{-1}N \geq 0$). So regular implies weak regular. The following result goes back to Varga [32].

LEMMA 3.1. *Let* $A \in \mathbf{R}^{n\times n}$ *be nonsingular and* $A = M - N$ *be a weak regular splitting. Then* $\rho(M^{-1}N) < 1$ *if and only if* $A^{-1} \geq 0$.

Here, $\rho(\cdot)$ denotes the spectral radius of a matrix in $\mathbf{R}^{n\times n}$. A special class of matrices A with $A^{-1} \geq 0$ is given by the M-matrices. A nonsingular matrix $A = (a_{ij}) \in \mathbf{R}^{n\times n}$ is termed an *M-matrix* if $a_{ij} \leq 0$ for $i \neq j$ and $A^{-1} \geq 0$. Alternatively, instead of $A^{-1} \geq 0$ we can equivalently require $Au > 0$ for some vector $u > 0$ (see [2]). The latter characterization shows immediately that the following lemma is valid [22, sect. 2.4.10].

LEMMA 3.2. *Let* $A = (a_{ij}), B = (b_{ij}) \in \mathbf{R}^{n\times n}$ *and assume that* A *is an* M*-matrix,* $b_{ij} \leq 0$ *for* $i \neq j$ *and* $A \leq B$. *Then* B *is an* M*-matrix.*

For a given matrix $A = (a_{ij}) \in \mathbf{R}^{n\times n}$, its *comparison matrix* $\langle A \rangle = (\alpha_{ij}) \in \mathbf{R}^{n\times n}$ is defined by

$$\alpha_{ij} = \begin{cases} |a_{ii}| & \text{if } i = j, \\ -|a_{ij}| & \text{if } i \neq j. \end{cases}$$

A is called an *H-matrix* if $\langle A \rangle$ is an M-matrix. The previous lemma immediately yields an analogous result for H-matrices (see [8], for example).

LEMMA 3.3. *Let* $A, B \in \mathbf{R}^{n\times n}$ *such that* $\langle A \rangle \leq \langle B \rangle$. *If* A *is an* H*-matrix, then* B *is an* H*-matrix as well.*

H-matrices are always nonsingular (see [2]). According to our previous remark on M-matrices, $A = (a_{ij})$ being an H-matrix is characterized by the existence of a positive vector u such that $\langle A \rangle u > 0$. Writing this componentwise yields

$$|a_{ii}|u_i > \sum_{j=1, j\neq i}^{n} |a_{ij}|u_j \ , \quad i = 1, \ldots, n \ .$$

Therefore, H-matrices may be viewed as generalized diagonally dominant matrices with weights u_i. Special H-matrices include strictly diagonally dominant matrices (take $u = (1, \ldots, 1)^T$) as well as irreducibly diagonally dominant matrices or weakly Ω-diagonally dominant matrices (see [7], [19]).

The absolute value $|A|$ of $A = (a_{ij}) \in \mathbf{R}^{n\times n}$ is again defined componentwise; i.e., $|A| = (|a_{ij}|) \in \mathbf{R}^{n\times n}$. Our final auxiliary result in this section gives a bound on $|A^{-1}|$ if A is an H-matrix (see [8] or [23]).

LEMMA 3.4. *Let $A \in \mathbf{R}^{n\times n}$ be an H-matrix. Then $|A^{-1}| \le \langle A\rangle^{-1}$.*

We finally define the weighted max-norm $||\cdot||_u$ in $\mathbf{R}^n$ by $||x||_u = \max_{1\le i\le n} |x_i/u_i|$, where $u \in \mathbf{R}^n$ is a positive vector. An easy calculation shows that the induced operator norm is given by

$$||A||_u = \max_{1\le i\le n} \left\{ \frac{1}{u_i} \sum_{j=1}^{n} \mid a_{ij} \mid u_j \right\} .$$

4. Convergence. According to Definition 2.1 we have L generally different iterates $x^{k,l}, l = 1, \ldots, L$, at each iteration k in a weighted additive Schwarz-type method. We collect them in the vector

$$\boldsymbol{x}^k = (x^{k,1}, \ldots, x^{k,L})^T \in \mathbf{R}^{Ln} .$$

Defining

$$\boldsymbol{c} = \left(\sum_{m=1}^{L} E_{1,m} M_m^{-1} b, \ldots, \sum_{m=1}^{L} E_{L,m} M_m^{-1} b \right)^T \in \mathbf{R}^{Ln}$$

and

$$\boldsymbol{H} = \begin{bmatrix} E_{1,1}M_1^{-1}N_1 & \cdots & E_{1,L}M_L^{-1}N_L \\ \vdots & \ddots & \vdots \\ E_{L,1}M_1^{-1}N_1 & \cdots & E_{L,L}M_L^{-1}N_L \end{bmatrix} \in \mathbf{R}^{Ln\times Ln}, \tag{14}$$

we can rewrite (7) and (8) as

$$\boldsymbol{x}^{k+1} = \boldsymbol{H}x^k + \boldsymbol{c} . \tag{15}$$

Our convergence analysis will be based on (15). We start by showing that (15) is consistent with $Ax = b$.

LEMMA 4.1. *Let $A \in \mathbf{R}^{n\times n}$ be nonsingular, $x^* = A^{-1}b$, and $\boldsymbol{x}^* = (x^*, \ldots, x^*)^T \in \mathbf{R}^{Ln}$. Then*

$$\boldsymbol{x}^* = \boldsymbol{H}\boldsymbol{x}^* + \boldsymbol{c} . \tag{16}$$

Consequently, if $\rho(\boldsymbol{H}) < 1$, then $\boldsymbol{x}^*$ is the unique fixed point of the affine operator $\boldsymbol{x} \mapsto \boldsymbol{H}\boldsymbol{x} + \boldsymbol{c}$.

Proof. Since $Ax^* = b$, we have $x^* = M_m^{-1}(N_m x^* + b)$ for $m = 1, \ldots, L$, and thus

$$\begin{aligned} \sum_{m=1}^{L} E_{l,m} M_m^{-1} N_m x^* + \sum_{m=1}^{L} E_{l,m} M_m^{-1} b &= \sum_{m=1}^{L} E_{l,m} M_m^{-1} (N_m x^* + b) \\ &= \sum_{m=1}^{L} E_{l,m} x^* = x^* \quad \text{for } l = 1, \ldots, L . \end{aligned}$$

This proves (16). If $\rho(\boldsymbol{H}) < 1$, the matrix $\boldsymbol{I} - \boldsymbol{H}$ is nonsingular, so $\boldsymbol{x} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{c} \Leftrightarrow (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{x} = \boldsymbol{c}$ has a unique solution, which is $\boldsymbol{x}^*$. ∎

To formulate our general convergence result we need the following basic auxiliary result (see also [32], for example).

LEMMA 4.2. *Let $A = (a_{ij}) \in \mathbf{R}^{n\times n}$ and suppose there exists $u \in \mathbf{R}^n$, $u > 0$, such that*

$$|A|u < u \ . \tag{17}$$

Then $\rho(A) < 1$. Moreover, let $\Theta \in [0, 1)$ be such that

$$|A|u \leq \Theta u \ . \tag{18}$$

Then $\rho(A) \leq \Theta < 1$.

Proof. We recall $||A||_u = \max_{i=1}^n \left(\sum_{j=1}^n \mid a_{ij} \mid u_j\right)/u_i$, so in the case of (17) we have $||A||_u < 1$, and in the case of (18) we have $||A||_u \leq \Theta$. Both assertions now follow, since $\rho(A) \leq ||A||$ for any operator norm. ∎

THEOREM 4.3. *Let a weighted additive Schwarz-type splitting of $A \in \mathbf{R}^{n\times n}$ with splittings $A = M_l - N_l$ and weighting matrices $E_{l,m}$, $l, m = 1, \ldots, L$, be given. Assume that there exists a vector $u \in \mathbf{R}^n, u > 0$, such that $|M_l^{-1}N_l|u < u, \quad l = 1, \ldots, L$. Then $\rho(\boldsymbol{H}) < 1$, which implies $\lim_{k\to\infty} x^{k,l} = x^* = A^{-1}b$, $l = 1, \ldots, L$, for the iterates of the corresponding iteration* (7), (8).

Proof. In light of Lemma 4.1, we only have to show that $\rho(\boldsymbol{H}) < 1$. Denote

$$\boldsymbol{u} = (u, \ldots, u)^T \in \mathbf{R}^{Ln} \ .$$

Obviously, $\boldsymbol{u} > 0$. The lth block of $|\boldsymbol{H}|\boldsymbol{u}$ is given by

$$\sum_{m=1}^{L} |E_{l,m}M_m^{-1}N_m|u = \sum_{m=1}^{L} E_{l,m}|M_m^{-1}N_m|u < \sum_{m=1}^{L} E_{l,m}u = u,$$

which results in $|\boldsymbol{H}|\boldsymbol{u} < \boldsymbol{u}$, so $\rho(\boldsymbol{H}) < 1$ follows from Lemma 4.2. ∎

In weighted additive Schwarz-type methods, generally, different iterates are computed on overlaps. The above theorem includes the important assertion that all sequences of different iterates on an overlap converge to the same solution. For special cases, this was already observed in [6], [25].

COROLLARY 4.4. *Assume that each splitting $M_l - N_l$, $l = 1, \ldots, L$, in a weighted additive Schwarz-type splitting of $A \in \mathbf{R}^{n\times n}$ is weak regular. Moreover, assume $A^{-1} \geq 0$. Then $\rho(\boldsymbol{H}) < 1$.*

Proof. Let $e = (1, \ldots, 1)^T \in \mathbf{R}^n$ and $u = A^{-1}e$. Since $A^{-1} \geq 0$, we have $u > 0$. Now $|M_l^{-1}N_l| = M_l^{-1}N_l = I - M_l^{-1}A$, and thus $|M_l^{-1}N_l|u = u - M_l^{-1}Au = u - M_l^{-1}e < u$, the last inequality holding because of $M_l^{-1} \geq 0$. The conclusion thus follows from Theorem 4.3. ∎

Another corollary arises for particular splittings of an H-matrix.

COROLLARY 4.5. *Let A be an H-matrix and assume that each splitting $M_l - N_l$ in a weighted additive Schwarz-type splitting satisfies*

$$\langle A\rangle \leq \langle M_l\rangle - |N_l|, \qquad l = 1, \ldots, L \ . \tag{19}$$

Then $\rho(\boldsymbol{H}) < 1$.

Proof. Because of (19) each $\langle M_l \rangle$ satisfies $\langle A \rangle \leq \langle M_l \rangle$. With Lemma 3.3, this implies that M_l is an H-matrix. Therefore $|M_l^{-1} N_l| \leq |M_l^{-1}|\,|N_l| \leq \langle M_l \rangle^{-1} |N_l|$, $l = 1, \ldots, L$, by Lemma 3.4. Together with (19) this implies $|M_l^{-1} N_l| \leq \langle M_l \rangle^{-1} (\langle M_l \rangle - \langle A \rangle) = I - \langle M_l \rangle^{-1} \langle A \rangle$, $l = 1, \ldots, L$. As in Corollary 4.4, let $e = (1, \ldots, 1)^T$ and $u = \langle A \rangle^{-1} e > 0$. Then $|M_l^{-1} N_l| u \leq u - \langle M_l \rangle^{-1} e < u$, $l = 1, \ldots, L$, so that we are back to the situation of Theorem 4.3. ∎

Special cases of Corollary 4.5 arise if

$$\langle A \rangle = \langle M_l \rangle - |N_l|, \qquad l = 1, \ldots, L\ . \tag{20}$$

Such splittings were termed *H-compatible* in [12]. The equality (20) holds trivially if each entry of M_l is either 0 or equal to the corresponding entry of A, as is the case in the traditional (block) Jacobi or (block) Gauß–Seidel splittings, for example. Also note that Corollary 4.5 holds, in particular, for M-matrices, since these are special cases of H-matrices.

5. Two-stage variants. The splittings $A = M_l - N_l, l = 1, \ldots, L$, are often primarily conceptual rather than numerically practical: they describe the principal way in which the original system $Ax = b$ is decomposed for the general additive Schwarz-type process, but systems with the matrices M_l can be too expensive to be solved exactly. Therefore, analogously to [11], [13], [15], [16], [25], [30], e.g., we now consider the situation where for each primary or *outer* splitting $A = M_l - N_l$, we have an additional *inner* splitting

$$M_l = F_l - G_l, \qquad l = 1, \ldots, L\ .$$

Instead of solving

$$M_l y = N_l x^{k,l} + b \tag{21}$$

for y to obtain $y^{k,l}$ in the weighted additive Schwarz-type method (7), (8), we now approximate the solution of (21) by performing s steps of the *inner iteration*

$$F_l y^{k,l,\nu+1} = G_l y^{k,l,\nu} + N_l x^{k,l} + b, \qquad \nu = 0, \ldots, s-1, \tag{22}$$

with $y^{k,l,0} = x^{k,l}$, taking $y^{k,l} = y^{k,l,s}$ in (7). Of course, the number s of inner iterations may depend on the iteration level k and on the individual outer splittings, so we write $s = s(k,l)$. A short calculation shows

$$\begin{aligned} y^{k,l} = y^{k,l,s(k,l)} &= (F_l^{-1} G_l)^{s(k,l)} x^{k,l} + \sum_{\nu=0}^{s(k,l)-1} (F_l^{-1} G_l)^\nu F_l^{-1} (N_l x^{k,l} + b) \\ &= \left(I - (I - (F_l^{-1} G_l)^{s(k,l)}) M_l^{-1} A \right) x^{k,l} + c^{s(k,l)} \end{aligned}$$

with

$$c^{s(k,l)} = \sum_{\nu=0}^{s(k,l)-1} (F_l^{-1} G_l)^\nu F_l^{-1} b\ . \tag{23}$$

Define

$$T_{s(k,l)} = (F_l^{-1} G_l)^{s(k,l)} + \sum_{\nu=0}^{s(k,l)-1} (F_l^{-1} G_l)^\nu F_l^{-1} N_l \tag{24}$$

$$= I - \left(I - (F_l^{-1} G_l)^{s(k,l)} \right) M_l^{-1} A \tag{25}$$

$$= (M_{s(k,l)})^{-1} N_{s(k,l)} \tag{26}$$

with

$$M_{s(k,l)} = M_l \left(I - (F_l^{-1} G_l)^{s(k,l)} \right)^{-1}, \quad N_{s(k,l)} = A - M_{s(k,l)} \ , \tag{27}$$

so $A = M_{s(k,l)} - N_{s(k,l)}$ is the unique splitting leading to $T_{s(k,l)}$; see [15]. Then the *two-stage weighted additive Schwarz-type method* is given by

$$x^{k+1,l} = \sum_{m=1}^{L} E_{l,m} y^{k,m}, \qquad k = 0, 1, \ldots, \tag{28}$$

where

$$y^{k,m} = T_{s(k,m)} x^{k,m} + c^{s(k,m)}, \qquad m = 1, \ldots, L,$$

with $T_{s(k,m)}$ from (24), $c^{s(k,m)}$ from (23), and $E_{l,m}$ from Definition 2.1.

As in section 4 we can write the whole process as an iteration in $\mathbf{R}^{n \times n}$ by setting

$$\boldsymbol{x}^k = (x^{k,1}, \ldots, x^{k,L}), \qquad \boldsymbol{c}^k = \left(\sum_{m=1}^{L} E_{1,m} c^{s(k,m)}, \ldots, \sum_{m=1}^{L} E_{L,m} c^{s(k,m)} \right)^T,$$

and

$$\boldsymbol{H}_k = \begin{pmatrix} E_{1,1} T_{s(k,1)} & \cdots & E_{1,L} T_{s(k,L)} \\ \vdots & & \vdots \\ E_{L,1} T_{s(k,1)} & \cdots & E_{L,L} T_{s(k,L)} \end{pmatrix}$$

so that

$$\boldsymbol{x}^{k+1} = \boldsymbol{H}_k \boldsymbol{x}^k + \boldsymbol{c}^k, \qquad k = 0, 1, \ldots. \tag{29}$$

Note that $x^* = A^{-1} b$ satisfies $F_l x^* = G_l x^* + N_l x^* + b$ for $l = 1, \ldots, L$, so from (22) we deduce that $x^* = T_{s(k,l)} x^* + c^{s(k,l)}$. Defining the error $\boldsymbol{e}^k$ of the kth iteration as $\boldsymbol{e}^k = \boldsymbol{x}^k - \boldsymbol{x}^*$, where $\boldsymbol{x}^* = (x^*, \ldots, x^*)^T \in \mathbf{R}^{Ln}$, we obtain

$$\boldsymbol{e}^{k+1} = \boldsymbol{H}_k \boldsymbol{e}^k, \qquad k = 0, 1, \ldots. \tag{30}$$

We are now ready to prove our main result on the convergence of the two-stage iteration (29). It is similar to Theorem 4.3, and we will subsequently again give two corollaries illustrating special cases of this theorem. Also note that the theorem and the corollaries hold for arbitrary $s(k,l)$, and so they hold particularly for the *nonstationary* case where $s(k,l)$ depends explicitly on k. Of course, the results also hold for the *stationary* case where $s(k,l) = s(l)$.

THEOREM 5.1. *Assume that there exists a positive vector $u \in \mathbf{R}^n, u > 0$, such that for $l = 1, \ldots, L$ and for all $k = 0, 1, \ldots$ we have*

$$|T_{s(k,l)}| u \leq \Theta u \tag{31}$$

with $\Theta \in [0, 1)$. Then the two-stage iterates $\boldsymbol{x}^k$ from (29) *satisfy $\lim_{k \to \infty} \boldsymbol{x}^k = \boldsymbol{x}^*$.*

Proof. By (30), the errors $\boldsymbol{e}^k$ satisfy

$$\boldsymbol{e}^k = \boldsymbol{H}_{k-1} \boldsymbol{e}^{k-1} = \boldsymbol{H}_{k-1} \cdots \boldsymbol{H}_0 \boldsymbol{e}^0, \qquad k = 1, 2, \ldots.$$

So we have to show that $\lim_{k\to\infty} \boldsymbol{H}_{k-1}\cdots\boldsymbol{H}_0 = 0$. Defining $\boldsymbol{u} = (u,\ldots,u)^T$, from (31) we obtain for the lth block in $|\boldsymbol{H}_k|\boldsymbol{u}$

$$\sum_{m=1}^{L} |E_{l,m}T_{s(k,m)}|u = \sum_{m=1}^{L} E_{l,m}|T_{s(km)}|u \le \sum_{m=1}^{L} E_{l,m}\Theta u = \Theta u \ ,$$

so $|\boldsymbol{H}_k|\boldsymbol{u} \le \Theta\boldsymbol{u}$ and, consequently, by Lemma 4.2

$$||\boldsymbol{H}_k||\boldsymbol{u} \le \Theta \ , \qquad k = 0,1,\ldots .$$

This implies $||\boldsymbol{H}_{k-1}\cdots\boldsymbol{H}_0||\boldsymbol{u} \le \Theta^k$ and thus $\lim_{k\to\infty} \boldsymbol{H}_{k-1}\cdots\boldsymbol{H}_0 = 0$. ∎

COROLLARY 5.2. *Suppose that the splittings $A = M_l - N_l$, $l = 1,\ldots,L$, and $M_l = F_l - G_l$, $l = 1,\ldots,L$, are all weak regular. Assume that $A^{-1} \ge 0$ and $F_l^{-1}N_l \ge 0$, $l = 1,\ldots,L$. Then $\lim_{k\to\infty} \boldsymbol{x}^k = \boldsymbol{x}^*$ for the two-stage iterates $\boldsymbol{x}^k$ from* (29).

Proof. Let $e = (1,\ldots,1)^T$ and $u = A^{-1}e > 0$. Since $F_l^{-1}G_l \ge 0, F_l^{-1} \ge 0$, and $F_l^{-1}N_l \ge 0$, we see that $T_{s(k,l)} \ge 0$ from (24). Moreover, since by Lemma 3.1 $\rho(F_l^{-1}G_l) < 1$, we have $M_l^{-1} = (I - F_l^{-1}G_l)^{-1}F_l^{-1} = \sum_{j=0}^{\infty}(F_l^{-1}G_l)^j F_l^{-1}$. Therefore, from the representation of $T_{s(k,l)}$ in (25), we get

$$\begin{aligned} 0 \ \le \ |T_{s(k,l)}|u = T_{s(k,l)}u &= u - \left(I - (F_l^{-1}G_l)^{s(k,l)}\right) M_l^{-1}e \\ &= u - \left(I - (F_l^{-1}G_l)^{s(k,l)}\right) \sum_{j=0}^{\infty}(F_l^{-1}G_l)^j F_l^{-1}e \\ &= u - \sum_{j=0}^{s(k,l)-1} (F_l^{-1}G_l)^j F_l^{-1}e \ \le \ u - F_l^{-1}e \ . \end{aligned}$$

Since $F_l^{-1}e > 0$, there exists $\Theta_l \in [0,1)$ such that $|T_{s(k,l)}|u \le \Theta_l u$, $l = 1,\ldots,L$. Taking $\Theta = \max_{1\le l\le L} \Theta_l$, we finally obtain $|T_{s(k,l)}|u \le \Theta u$ with $\Theta \in [0,1)$, as required for Theorem 5.1. ∎

In passing, let us note that Corollary 5.2 generalizes results for single splitting two-stage methods in [12], [16], where the outer splitting was assumed to be regular and the inner to be weak regular.

COROLLARY 5.3. *Let A be an H-matrix and assume that all splittings $A = M_l - N_l$ satisfy*

$$\langle A\rangle \le \langle M_l\rangle - |N_l|, \qquad l = 1,\ldots,L, \tag{32}$$

and that all inner splittings $M_l = F_l - G_l$ satisfy

$$\langle M_l\rangle = \langle F_l\rangle - |G_l|, \qquad l = 1,\ldots,L \ . \tag{33}$$

Then $\lim_{k\to\infty} \boldsymbol{x}^k = \boldsymbol{x}^*$ for the two-stage iterates $\boldsymbol{x}^k$ from (29).

Proof. The proof follows by combining the techniques in the proofs of Corollaries 5.2 and 4.5. ∎

6. Comparison theorems. So far we have proved the convergence of weighted additive Schwarz-type methods. A natural question that arises in this context is, for example, how the size of the overlap influences the rate of convergence. Any theoretical investigation in this direction seems to be very difficult even for qualitative assertions.

For multisplittings, some results are found in [1], [10], but their technique of proof does not carry over to our more general case.

We will give a result comparing a weighted additive Schwarz-type splitting with a single splitting. It is similar to those given in [5], [20] for multisplittings. We need the following auxiliary result.

LEMMA 6.1. *Let $A \in \mathbf{R}^{n\times n}$ be nonnegative, $u \in \mathbf{R}^n$ be positive, and $\Theta > 0$.*

(i) *If $Au \le \Theta u$, then $\rho(A) \le \Theta$.*

(ii) *If $Au \ge \Theta u$, then $\rho(A) \ge \Theta$.*

Proof. (i) was proved in Lemma 4.2. For (ii) we have

$$0 < u \le \left(\frac{1}{\Theta}A\right) u \le \cdots \le \left(\frac{1}{\Theta}A\right)^k u \le \cdots,$$

which shows that $\left(\frac{1}{\Theta}A\right)^k$ cannot converge to 0. This implies $\rho\left(\frac{1}{\Theta}A\right) \ge 1$, i.e., $\rho(A) \ge \Theta$. ∎

THEOREM 6.2. *Let $A \in \mathbf{R}^{n\times n}$ be nonsingular with $A^{-1} \ge 0$. Let $A = M - N = \widetilde{M} - \widetilde{N}$ be two regular splittings of A. Moreover, assume that in a weighted additive Schwarz-type splitting of A, all splittings $A = M_l - N_l$ are weak regular and*

$$\widetilde{M}^{-1} \ge M_l^{-1} \ge M^{-1}, \qquad l = 1, \ldots, L\ .$$

Let $\boldsymbol{H}$ denote the iteration matrix for the weighted additive Schwarz-type splitting as defined in (14). *Then*

$$\rho(\widetilde{M}^{-1}\widetilde{N}) \le \rho(\boldsymbol{H}) \le \rho(M^{-1}N).$$

Proof. We start proving the second inequality. For simplicity, let us write ρ instead of $\rho(M^{-1}N)$. Note that $\rho < 1$ due to Lemma 3.1. By the Perron–Frobenius theorem (see [2, 24, 31]) there exists $u \in \mathbf{R}^n$, $u \ge 0$, such that $M^{-1}Nu = \rho u$. Since $0 \le Nu = \rho Mu$ we get $Au = Mu - Nu = (1-\rho)Mu \ge 0$. Thus, for each $l = 1, \ldots, L$, we have

$$M_l^{-1}N_l u = (I - M_l^{-1}A)u = u - M_l^{-1}Au \le u - M^{-1}Au = M^{-1}Nu = \rho u\ .$$

Defining $\boldsymbol{u} = (u, \ldots, u) \in \mathbf{R}^{Ln}$, this immediately yields $\boldsymbol{H}\boldsymbol{u} \le \rho\boldsymbol{u}$ with $\boldsymbol{H} \ge 0$. Hence, if u is positive, $\rho(\boldsymbol{H}) \le \rho$ by Lemma 6.1 (i). If u is only nonnegative, define

$$E = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \in \mathbf{R}^{n\times n}$$

and consider the matrix $A_\epsilon = A - \epsilon E$ with the splittings $A_\epsilon = M - (N + \epsilon E)$, $A_\epsilon = M_l - (N_l + \epsilon E)$, $l = 1, \ldots, L$. For all $\epsilon > 0$ the matrix $M^{-1}(N + \epsilon E)$ is positive so that there exists a *positive* Perron vector u_ϵ with $M^{-1}(N + \epsilon E)u_\epsilon = \rho_\epsilon u_\epsilon$, $\rho_\epsilon := \rho(M^{-1}(N + \epsilon E)) > 0$. Taking $\epsilon > 0$ sufficiently small, we still have $\rho_\epsilon < 1$, and thus $A_\epsilon^{-1} = \left(\sum_{\nu=0}^{\infty}(M^{-1}(N + \epsilon E))^\nu\right) M^{-1} \ge 0$. Repeating the part of the proof given before, we thus get $A_\epsilon u_\epsilon \ge 0$ and $\rho(\boldsymbol{H}_\epsilon) \le \rho_\epsilon$ for the iteration matrix of the corresponding weighted additive Schwarz-type splitting. As ϵ tends to 0 this yields $\rho(\boldsymbol{H}) \le \rho$.

The proof for the first inequality proceeds in a completely analogous manner showing $\boldsymbol{H}\widetilde{\boldsymbol{u}} \geq \rho(\widetilde{M}^{-1}\widetilde{N})\widetilde{\boldsymbol{u}}$ for the Perron vector $\widetilde{u}$ of $\widetilde{M}^{-1}\widetilde{N}$. If $\widetilde{u} > 0$ the proof is complete according to Lemma 6.1 (ii); otherwise we consider A_ϵ, as before. ∎

To give a specific example, let $A = D - L - U$ be an M-matrix with diagonal part D and strictly lower and upper triangular parts $-L$ and $-U$, respectively. Let $S_1, \ldots, S_L$ be an (overlapping) decomposition of $\{1, \ldots, n\}$ and define

$$(M_l)_{ij} = \begin{cases} (A)_{ij} & \text{if } (i = j) \text{ or } (i > j \text{ and } i, j \in S_l), \\ 0 & \text{otherwise.} \end{cases} \tag{34}$$

With $M = D$, $\widetilde{M} = D - L$, all assumptions of Theorem 6.2 are met, showing that any weighted additive Schwarz-type splitting with the matrices M_l from (34) converges at least as fast as the Jacobi iteration but not faster than the Gauß–Seidel method.

As another particular case, assume that all splittings are equal; i.e., $M = \widetilde{M} = M_l$, $l = 1, \ldots, L$. Theorem 6.2 then shows that $\rho(\boldsymbol{H}) = \rho(M^{-1}N)$. This equality even holds under less restrictive assumptions, as the following theorem shows.

THEOREM 6.3. *Assume a splitting $A = M - N$ and let $M_l = M, N_l = N, l = 1, \ldots, L$, in (8). Assume further that $H \geq 0$, $H \neq 0$ for $H = M^{-1}N$. Then for any weighted additive Schwarz-type splitting,*

$$\rho(H) = \rho(\boldsymbol{H}),$$

where $\boldsymbol{H}$ denotes the iteration matrix from (14).

Proof. Let $\epsilon > 0$ such that

$$H_\epsilon = H + \epsilon E$$

has a positive Perron vector $u \in \mathbf{R}^n$. An immediate calculation shows that for $\boldsymbol{u} = (u, \ldots, u) \in \mathbf{R}^{Ln}$ we have $\boldsymbol{H}_\epsilon \boldsymbol{u} = \rho(H_\epsilon)\boldsymbol{u}$ so that Lemma 6.1(i),(ii) yields $\rho(\boldsymbol{H}_\epsilon) = \rho(H_\epsilon)$ and, as $\epsilon \to 0$, $\rho(\boldsymbol{H}) = \rho(H)$. ∎

The above theorem can be adapted to two-stage methods as the subsequent corollary shows.

COROLLARY 6.4. *Assume a stationary two-stage splitting $A = M - N, M = F - G$ and let $M_l = M, N_l = N, F_l = F, G_l = G, l = 1, \ldots, L$, in (24) with $s(k, l) = s$ for all l, k. Assume further that $A = M - N, M = F - G$ are weak regular splittings and that $F^{-1}N \geq 0$.*

Then

$$\rho(T_s) = \rho(\boldsymbol{H}_s),$$

where $\boldsymbol{H}_s$ denotes the iteration matrix (14) for this stationary two-stage method.

Proof. By (26), (27) we have $T_s = M_s^{-1}N_s$ with $M_s = M(I - (F^{-1}G)^s)^{-1}, N_s = A - M_s$. The inequalities $F^{-1}N, F^{-1}, F^{-1}G \geq 0$ yield $T_s \geq 0$. $M^{-1}, F^{-1}, F^{-1}G \geq 0$ imply $\rho(F^{-1}G) < 1$, hence $(I - (F^{-1}G)^s)^{-1}$ exists and M_s is well defined. We further note $M_s^{-1} = (I - (F^{-1}G)^s)M^{-1} = \sum_{j=0}^{s-1}(F^{-1}G)^j F^{-1} \geq 0$. We now repeat the proof of Theorem 6.3 with M, N replaced by M_s, N_s. ∎

The above theorem and its corollary generalize results to both one-stage and two-stage variants of weighted additive Schwarz-type methods, which have been proved in [25], [27] for one-stage Schwarz-type methods (more precisely: two-stage methods with one inner iteration) and more abstract results presented in [20] for one-stage multisplitting methods. The above theorem suggests, however, that all these results are

of more theoretical than practical interest: if we start the weighted additive Schwarz-type iteration with identical initial guesses $x^{0,1} = \cdots = x^{0,L} = x^0$, we get (see (14)) identical iterates $x^{k,1} = \cdots = x^{k,L} = x^k$, where x^k is also obtained by the single splitting iteration $x^{k+1} = M^{-1}(Nx^k + b)$, $k = 0, 1, \ldots$, with initial guess x^0.

7. Algorithmic aspects. For multisplitting methods, implementation details are described in [3, Chap. 7, 9]. Here, we discuss some algorithmic aspects of general weighted additive Schwarz-type methods. We assume L processors $P(m)$, $m = 1, \ldots, L$, each associated with one splitting $A = M_m - N_m$. For simplicity we do not include the convergence control.

In the most general case following Definition 2.1, iteration k can be implemented as follows in a message passing environment:

for $m = 1$ **to** L **do in parallel on** $P(m)$
  solve $M_m y = N_m x^{k,m} + b$ for y exactly or approximately (in a two-stage method),
    call result $y^{k,m}$
  for $l = 1$ **to** $L, l \neq m$ **do**
    send $E_{l,m} y^{k,m}$ to $P(l)$
    receive $E_{m,l} y^{k,l}$ from $P(l)$
  accumulate $x^{k+1,m} = \sum_{l=1}^{L} E_{m,l} y^{k,l}$

In the above algorithm, each processor $P(m)$ holds a complete iterate $x^{k,m}$. In every step, the processors exchange "their" respective approximations $y^{k,m}$ (which, for convenience, are already multiplied by the weighting factors $E_{l,m}$ required by the receiving processor). These are then accumulated to yield the next complete iterate $x^{k+1,m}$.

The algorithmic efficiency of communication is determined by the *message length*, which is formally equal to n for each message $E_{l,m} y^{k,m}$, and the maximal *number of messages*, which is $2L(L-1)$ in the above general algorithm. Both can be substantially reduced in specific situations.

The message *length* can be reduced as soon as the $E_{l,m}$ have many zero diagonal elements. It is then advantageous to restrict the messages to the nonzero components of $E_{l,m} y^{k,m}$. To be specific, define $S_m = \{i : 1 \leq i \leq n \text{ and } (E_{m,m})_{ii} \neq 0\}$. The block property of Definition 2.4 then ensures that each processor $P(m)$ only needs to compute those components of $y^{k,m}$ which belong to S_m. Due to the block property, only a part of these nonzero components must be sent to the other processors since $(E_{l,m} y^{k,m})_i = 0$ as soon as $i \notin S_m$. Hence, when transferring $E_{lm} y^{k,m}$, the message length actually reduces to at most $|S_m|$ components.

Furthermore, in applications with a sparse matrix A (and sparse matrices M_m, N_m), the *number* of messages will usually also be reduced drastically. Typically then, a processor only communicates with a few "neighbors," since each processor $P(m)$ needs only a small part of $x^{k,m}$ to form those components of $N_m(x^{k,m} + b)$ required to compute its part of $y^{k,m}$. This part of $x^{k,m}$ will usually involve only a few of the $E_{m,l} y^{k-1,l}, l \neq m$, the nonzero components of which are the only ones to be communicated.

As opposed to methods which perform a "true" weighting of components, additive Schwarz methods according to Definition 2.3 tend to minimize the overall message length. Each processor receives only one value for each relevant component of $x^{k,m+1}$, and the accumulation to get $x^{k+1,m}$ reduces to a mere selection process. If a "true" weighting is performed for some component, more than one value for this component has to be transferred and the weighting requires arithmetic work during the

accumulation process. This shows that methods with a "true" weighting exhibit a certain overhead with respect to computation and communication as compared to the additive Schwarz methods from Definition 2.3.

In conclusion, we point out that the above algorithmic framework illustrates that the similarity of algebraic Schwarz methods and multisplitting methods exceeds the formal aspect of Definition 2.1. In the framework of overlapping block decompositions these methods simply differ by their treatment of the approximations on the overlaps. Schwarz methods keep different approximations on the overlaps during the whole iterative process, while multisplitting methods determine a unique approximation after every iteration. This requires an additional weighting process, producing some communicational and computational overhead.

8. Numerical results. For our numerical experiments we choose as an example the Dirichlet problem

$$(a(x)u_x)_x + (b(y)u_y)_y + \alpha u = 0$$

on a rectangle $\Omega = [0,1] \times [0,c]$ for $\alpha \in \{0.1, 1.0\}$ and

$$\begin{array}{lll} \text{(i)} & a(x) = 1 + 0.02x, & b(y) = 1 + 0.002y, \\ \text{(ii)} & a(x) = 1 + 2x, & b(y) = 0.133 + 1.2y, \end{array}$$

with Dirichlet boundary conditions

$$u(x,y) = \begin{cases} y & \text{for } x = 0, \\ 2y & \text{for } x = 1, \\ x/2 & \text{for } y = c, \\ 1.1x & \text{for } y = 0. \end{cases}$$

This example is a typical one, but it cannot be representative due to the large variety of possible applications. Nevertheless, it should give a feeling for the potential properties of the methods treated in this paper and, in particular, for the intended comparison.

The five-point discretization with central differences and a mesh size of $h = 1/(p+1)$ leads for $c = (q+1) \cdot h$ to the $q \times q$ block tridiagonal coefficient matrix

$$A = (-B_{j-1}, A_j, -B_j)_{j=1}^{q}$$

with diagonal $p \times p$ blocks

$$B_j = b\left(\tfrac{2j+1}{2}h\right) \cdot I$$

and tridiagonal $p \times p$ blocks

$$A_j = \left(-a(\tfrac{2i-1}{2}h), a(\tfrac{2i-1}{2}h) + a(\tfrac{2i+1}{2}h) + b(\tfrac{2j-1}{2}h) + b(\tfrac{2j+1}{2}h) + \alpha, -a(\tfrac{2i+1}{2}h)\right)_{i=1}^{p}.$$

The condition of A decreases as $\alpha > 0$ gets larger, so in all methods tested we expect $\alpha = 0.1$ to require more iterations than $\alpha = 1.0$. We introduce a simple domain decomposition of Ω into L overlapping rectangles

$$\begin{array}{ll} \Gamma_l = [0,1] \times [l(i), r(i)], & l = 1, \ldots, L, \\ l(i) = (i-1) \cdot q/L - ov, & r(i) = i \cdot q/L + 1 + ov, \\ l(1) = 0, & r(L) = q + 1. \end{array}$$

Each rectangle consists of $r(i) - l(i) - 1$ grid lines, each with two additional artificial boundaries $l(i), r(i)$ except for the first left and last right "true" boundaries. $2ov$ indicates the number of grid lines per overlap.

We report tests for a two-stage method (21), (22) with a block Jacobi iteration as outer iteration by choosing for $m = 1, \ldots, q$ the matrix M_m to be the identity apart from the blocks $l(m) + 1$ through $r(m) - 1$ where we take the corresponding entries of A; i.e.,

$$M_m = \begin{pmatrix} I & & & & & \\ & A_{l(m)+1} & -B_{l(m)+1} & & & \\ & -B_{l(m)+1} & \ddots & \ddots & & \\ & & \ddots & \ddots & -B_{r(m)-1} & \\ & & & -B_{r(m)-2} & A_{r(m)-1} & \\ & & & & & I \end{pmatrix}.$$

The inner iteration is also taken to be a block Jacobi iteration where for $m = 1, \ldots, q$ we remove the off-diagonal blocks in M_m; i.e.,

$$F_m = \begin{pmatrix} I & & & \\ & A_{l(m)+1} & & \\ & & \ddots & \\ & & & A_{r(m)-1} \\ & & & & I \end{pmatrix}.$$

We tested the following three variants:

schw	an algebraic additive Schwarz method,
ms-w	a multisplitting method with weight 0.5 on the overlaps,
ms-l	a multisplitting method, where on every overlap the last approximation of the upper neighboring subdomain is chosen after each step.

In *schw* we take

$$E_{m,m} = \begin{pmatrix} 0 & & \\ & I & \\ & & 0 \end{pmatrix},$$

whereas *ms-w* and *ms-l* are characterized by

$$E_m = \begin{pmatrix} 0 & & & \\ & \frac{1}{2}I & & \\ & & I & \\ & & & \frac{1}{2}I \\ & & & & 0 \end{pmatrix} \quad \text{and} \quad E_m = \begin{pmatrix} 0 & & & \\ & I & & \\ & & I & \\ & & & 0 \\ & & & & 0 \end{pmatrix},$$

respectively. Here, the horizontal and vertical lines refer to the same block decomposition as in M_m, F_m. Within the central block in the E_m, the first (last) diagonal block corresponds to that part which overlaps with the lower (upper) neighboring subdomain.

As a convergence criterion we use a modified relative difference of two iterates:

$$\max_{1\le i\le p,1\le j\le q}\left\{\frac{\left|x_{i,j}^{k+1}-x_{i,j}^{k}\right|}{\max\left\{\left|x_{i,j}^{k}\right|,10^{-300}\right\}}\right\}<10^{-14}.$$

In all tests with a given solution, we observed only minimal differences between the three variants with respect to the relative error.

In Table 1 we compare the three variants applied to example (ii) with $p=256, q=256, \alpha=0.1$, $L=16$ subdomains and $2ov=32$ overlapping lines. These (sequential) results have been obtained on a workstation IBM RS 6000-550 at the Department of Mathematics of the Technical University of Berlin. We state the number *itout* of outer iterations and the CPU time in seconds as a function of the number *itin* of inner iterations.

TABLE 1
Dependence on the number of inner iterations.

itin	*itout* *schw*	*itout* *ms-w*	*itout* *ms-l*	Time *schw*	Time *ms-w*	Time *ms-l*
1	1325	1325	1325	582.9	650.1	600.1
2	679	683	679	448.2	476.8	445.8
3	460	465	466	395.2	421.5	404.9
4	348	354	350	371.4	397.7	382.7
5	281	288	288	362.4	382.5	370.9
7	204	210	210	347.9	370.9	360.9
9	160	166	166	347.9	369.5	357.2
11	132	138	138	339.9	367.6	363.3
13	113	119	118	330.7	357.7	347.2
15	98	104	103	335.0	360.9	353.3
17	87	93	92	337.5	365.0	356.4

As one could expect, the number of outer iterations decreases with an increasing number of inner iterations. Since performing more inner iterations requires more time, there is an optimal number of inner iterations minimizing the CPU time. In the case of Table 1, $itin=13$ is optimal. In many other examples, only a few inner iterations (two to four) were optimal. As a characteristic result we obtain that for *itin* fixed the numbers of iterations needed by the three different methods vary only slightly, even for larger overlaps. In the latter case, however, the additional time for copying (sending) the whole overlaps in both multisplitting methods *ms-l* and *ms-w* and, in particular, that for the weighting process for method *ms-w*, is no longer negligible, so the algebraic additive Schwarz method becomes fastest. In particular, this effect becomes obvious for $itin=1$, where communication is necessary after every inner (= outer) iteration.

In Table 2 we illustrate the influence of the size of the overlaps. We take the above example for $itin=4, \alpha=0.1$, and $\alpha=1.0$. As the number of outer iterations differs by at most one or two for the three methods, we only note the results (computed again on an IBM RS 6000-550) for the algebraic additive Schwarz method.

Table 2 illustrates a typical behavior of weighted algebraic Schwarz-type methods. Starting with $ov=0$, i.e., a simple nonoverlapping block Jacobi method, the number of iterative steps decreases for a moderate increase in the size of the overlaps. For larger values of *ov*, the number of steps remains constant[1] and the increasing arithmetic

[1] As was suggested by one of the referees, a heuristic explanation for this fact is the very local character of the five-point stencil used in our discretization.

TABLE 2
Dependence on the size of the overlap.

$2ov$	*itout* ($\alpha = 0.1$)	Time ($\alpha = 0.1$)	*itout* ($\alpha = 1.0$)	Time ($\alpha = 1.0$)
0	384	180.7	61	28.6
2	356	175.4	46	22.6
4	353	182.4	45	23.2
6	351	188.5	44	24.2
8	351	197.5	44	26.2
10	350	207.6	44	26.4
12	350	219.2	44	28.3
14	349	226.9	44	28.7
16	349	236.5	44	29.8
18	349	245.2	44	31.2
20	349	256.6	44	33.0

complexity of the individual steps causes a re-increasing CPU time. In other cases the influence of the overlap can be stronger. See [10], for example, for results for a one-stage multisplitting method in the framework of waveform relaxation methods.

Our tests on a parallel machine have been carried out on the CRAY T3D of the Konrad-Zuse-Zentrum für Informationstechnik, Berlin. We used a message passing code for parallelization, which we implemented with CRAY PVM, a version particularly adapted to the hardware of the T3D.

Table 3 shows the scaling effect of the three methods for example (i) with $\alpha = 0.1, p = 500, itin = 4, ov = 1$, and $q = 2^{k+7}, 0 \le k \le 8$, the number of subdomains L increasing with and being equal to the number of processors.

TABLE 3
Scaling of the weighted additive Schwarz methods schw*,* ms-w*, and* ms-l.

q	# of proc	*itout schw*	*itout ms-w*	*itout ms-l*	Time *schw*	Time *ms-w*	Time *ms-l*
128	1	158	158	158	68.1	68.1	68.1
256	2	162	162	161	70.7	71.1	70.3
512	4	162	162	161	71.4	71.8	70.9
1024	8	163	162	161	71.9	71.9	71.5
2048	16	163	163	163	71.9	72.4	71.5
4096	32	164	164	163	72.5	72.9	72.1
8192	64	167	166	166	74.0	73.7	73.3
16384	128	171	171	171	78.0	76.9	76.4
32768	256	181	181	180	85.6	84.6	83.9

All three algorithms scale rather well. The number of outer iterative steps and the CPU time remain almost constant with an increasing number of processors (i.e., subdomains) up to 64 processors. For larger numbers of processors, the times begin to increase somewhat faster since the number of iterations increases (whereas the time per iterative step increases only very slowly). Note that the size of the subdomains $500 \times (128 + 2)$ is constant for $L > 2$, whereas for $L = 1$ the size is equal to 500×128 due to the lack of overlap.

Finally, our last table, Table 4, deals with the same example with $\alpha = 1.0, p = 64, q = 16\,384, itin = 4, ov = 1$. It illustrates the parallelization effect for a fixed problem size and an increasing number L of processors, i.e., an increasing number of subdomains with a decreasing size $p \times (q/L + 2)$. Since, again, the three methods

differ only minimally due to the small overlap, we only note the results for the Schwarz method *schw*. The case $L = 1$ could not be computed because of memory restrictions.

TABLE 4
Parallelization effect.

# of proc	$q/L + 2$	*itout*	Time	Speedup T_2/T_{proc}
2	8194	29	98.93	1.00
4	4098	30	50.07	1.98
8	2050	31	25.95	3.81
16	1026	31	12.94	7.65
32	514	31	6.99	14.15
64	258	32	3.43	28.85
128	130	32	2.01	49.02
256	66	32	1.86	53.18

For small numbers of processors, the difference to the optimal speedup is mostly due to the slightly different numbers of iterative steps. For large numbers of processors the speedup deviates somewhat more from its optimum due to the interference of the (not entirely proportionally) decreasing size of the subproblems and the increasing significance of communication. The latter effect becomes dominant for 256 processors and more.

In conclusion, the above numerical examples illustrate the strong similarity of algebraic additive Schwarz-type and multisplitting methods, as predicted by the theoretical results. The results do not suggest a clear preference when comparing these methods. A slight preference should probably be given to Schwarz-type methods in applications with more complicated communication patterns (like a three-dimensional problem where domain decomposition is applied in all three directions) and where a small number of inner iterations is optimal. In these cases the longer message length and the additional weighting process of a multisplitting method may increase the time.

Acknowledgments. We want to thank Daniel B. Szyld for many valuable comments and his very careful reading of the manuscript. We are also grateful to an anonymous referee for suggesting several improvements.

REFERENCES

[1] G. ALEFELD, I. LEHNHARDT, AND G. MAYER, *On multisplitting methods for band matrices*, Numer. Math., 75 (1997), pp. 267–292.
[2] A. BERMAN AND R. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979; reprinted by SIAM, Philadelphia, PA, 1994.
[3] K. BURRAGE, *Parallel and Sequential Methods for Ordinary Differential Equations*, Clarendon Press, Oxford, UK, 1995.
[4] T. CHAN AND T. MATHEW, *Domain decomposition algorithms*, Acta Numerica, (1994), pp. 61–144.
[5] L. ELSNER, *Comparison of weak regular splittings and multisplitting methods*, Numer. Math., 56 (1989), pp. 283–289.
[6] D. EVANS, S. JIANPING, AND K. LISHAN, *The convergence factor of the parallel Schwarz overrelaxation method for linear systems*, Parallel Comput., 6 (1988), pp. 313–324.
[7] A. FROMMER, *Generalized nonlinear diagonal dominance and applications to asynchronous iterative methods*, J. Comput. Appl. Math., 38 (1991), pp. 105–124.
[8] A. FROMMER AND G. MAYER, *Convergence of relaxed parallel multisplitting methods*, Linear Algebra Appl., 119 (1989), pp. 141–152.
[9] A. FROMMER AND G. MAYER, *On the theory and practice of multisplitting methods in parallel computation*, Computing, 49 (1992), pp. 63–74.

[10] A. Frommer and B. Pohl, *A comparison result for multisplittings and waveform relaxation methods*, Numer. Linear Algebra Appl., 2 (1995), pp. 335–346.
[11] A. Frommer and D. Szyld, *H-splittings and two-stage iterative methods*, Numer. Math., 63 (1992), pp. 65–82.
[12] A. Frommer and D. Szyld, *On asynchronous two-stage iterative methods*, Numer. Math., 69 (1994), pp. 141–153.
[13] G. Golub and M. Overton, *The convergence of inexact Chebyshev and Richardson iterative methods for solving linear systems of equations*, Numer. Math., 53 (1988), pp. 571–593.
[14] M. Griebel and P. Oswald, *Remarks on the abstract theory of additive and multiplicative Schwarz algorithms*, Numer. Math., 70 (1995), pp. 163–180.
[15] M. Jones and D. Szyld, *Two-stage multisplitting methods with overlapping blocks*, Numer. Linear Algebra Appl., 3 (1996), pp. 113–124.
[16] P. Lanzkron, D. Rose, and D. Szyld, *Convergence of nested classical iterative methods for linear systems*, Numer. Math., 58 (1991), pp. 685–702.
[17] P.-L. Lions, *On the Schwarz alternating method* I, in Proc. 1st International Symposium on Domain Decomposition Methods for Partial Differential Equations, Paris, January 1987, R. Glowinski, G. H. Golub, G. Meurant, and J. Périaux, eds., SIAM, Philadelphia, PA, 1988, pp. 1–42.
[18] P.-L. Lions, *On the Schwarz alternating method* II: *Stochastic interpretation and order properties*, in Domain Decomposition Methods - Proc. 2nd International Symposium on Domain Decomposition Methods for Partial Differential Equations, Los Angeles, January 1988, T. C. Chan, R. Glowinski, J. Périaux, and O. Widlund, eds., SIAM, Philadelphia, PA,1989, pp. 47–71.
[19] J. Moré, *Nonlinear generalizations of matrix diagonal dominance with application to Gauss–Seidel iterations*, SIAM J. Numer. Anal., 9 (1972), pp. 357–378.
[20] M. Neumann and R. Plemmons, *Convergence of parallel multisplitting iterative methods for M-matrices*, Linear Algebra Appl., 88/89 (1987), pp. 559–573.
[21] D. O'Leary and R. White, *Multisplittings of matrices and parallel solution of linear systems*, SIAM J. Algebraic Discrete Meth., 6 (1985), pp. 630–640.
[22] J. Ortega and W. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[23] A. Ostrowski, *Über die Determinanten mit überwiegender Hauptdiagonale*, Coment. Math. Helv., 10 (1937), pp. 69–96.
[24] W. Rheinboldt and J. Vandergraft, *A simple approach to the Perron-Frobenius theory for positive operators in general partially-ordered spaces*, Math. Comput., 27 (1973), pp. 139–145.
[25] G. Rodrigue, *Inner/outer iterative methods and numerical Schwarz algorithms*, Parallel Comput., 2 (1985), pp. 205–218.
[26] G. Rodrigue and J. Simon, *A generalized numerical Schwarz algorithm*, in Computer Methods in Applied Sciences and Engineering VI, R. Glowinski and J.-L. Lions, eds., Elsevier, New York, 1984, pp. 272–281.
[27] G. Rodrigue, K. LiShan, and L. Yu-Hui, *Convergence and comparison analysis of some numerical Schwarz methods*, Numer. Math., 56 (1989), pp. 123–138.
[28] B. Smith, P. Bjørstad, and W. Gropp, *Domain decomposition*, Cambridge University Press, Cambridge, UK, 1996.
[29] H. Schwarz, *Gesammelte mathematische Abhandlungen*, Vol. 2, 1890, Springer-Verlag, Berlin, pp. 133–134.
[30] D. Szyld and M. Jones, *Two-stage and multisplitting methods for the solution of linear systems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 671–679.
[31] J. Vandergraft, *Applications of partial orderings to the study of positive definiteness, monotonicity, and convergence of iterative methods for linear systems*, SIAM J. Numer. Anal., 9 (1972), pp. 97–104.
[32] R. Varga, *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1962.
[33] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 4, pp. 913–937, October 1997

009

ESTIMATING AN EIGENVECTOR BY THE POWER METHOD WITH A RANDOM START*

GIANNA M. DEL CORSO†

Abstract. This paper addresses the problem of approximating an eigenvector belonging to the largest eigenvalue of a symmetric positive definite matrix by the power method. We assume that the starting vector is randomly chosen with uniform distribution over the unit sphere.

This paper provides lower and upper as well as asymptotic bounds on the randomized error in the $\mathcal{L}_p$ sense, $p \in [1, +\infty]$. We prove that it is impossible to achieve sharp bounds that are independent of the ratio between the two largest eigenvalues. This should be contrasted to the problem of approximating the largest eigenvalue, for which Kuczyński and Woźniakowski [*SIAM J. Matrix Anal. Appl.*, 13 (1992), pp. 1094–1122] proved that it is possible to bound the randomized error at the kth step with a quantity that depends only on k and on the size of the matrix.

We prove that the rate of convergence depends on the ratio of the two largest eigenvalues, on their multiplicities, and on the particular norm. The rate of convergence is at most linear in the ratio of the two largest eigenvalues.

Key words. eigenvectors, power method, random start, randomized error

AMS subject classification. 65F15

PII. S0895479895296689

1. Introduction.

In this paper we deal with the power method, which is used to approximate a largest eigenvector of an $n \times n$ symmetric matrix A. By the largest eigenvector we mean a normalized eigenvector corresponding to the largest eigenvalue of A. Our analysis holds for every matrix A for which the power method is convergent. To simplify the analysis, we assume that A is positive definite.

It is well known that the convergence of the power method depends on the starting vector $\mathbf{b}$. In particular, the power method is not convergent if $\mathbf{b}$ is orthogonal to the eigenspace corresponding to the largest eigenvalue of A. Since no a priori information about this eigenspace is generally available, a random starting vector is usually chosen. This indicates the need to study the convergence of the power method with a random start.

It is easy to see that if $\mathbf{b}$ is randomly chosen according to the uniform distribution then the power method approximates a largest eigenvector and the largest eigenvalue with probability 1. The problem of approximating the largest eigenvalue by the power method with a random start has been considered in [4], where sharp upper bounds on the randomized relative error at each step are given. An important feature of these bounds is that they are independent of the distribution of the eigenvalues.

The approach of our paper is similar to that of [4]. We analyze the convergence of the power method for approximating a largest eigenvector when the starting vector $\mathbf{b}$ is randomly chosen with uniform distribution over the unit sphere of the n-dimensional space.

*Received by the editors December 22, 1995; accepted for publication (in revised form) by P. Van Dooren October 15, 1996. This work was done while the author was visiting the Computer Science Department at Columbia University. This research was (partially) supported by the ESPRIT III Basic Research Programme of the EC under contract 9072 (Project GEPPCOM).

http://www.siam.org/journals/simax/18-4/29668.html

†Dipartimento di Matematica, Università di Milano and IMC-CNR, Via Santa Maria 46, 56126 Pisa, Italy (delcorso@imc.pi.cnr.it).

In order to define the randomized error, we consider the acute angle $\alpha_k = \alpha_k(\mathbf{b})$ between the vector computed by the power method at the kth step and the eigenspace corresponding to the largest eigenvalue, and we study the expectation of $\sin(\alpha_k(\mathbf{b}))$ over $\mathbf{b}$ in the $\mathcal{L}_p$ sense, $p \in [1, +\infty]$.

We first ask whether it is possible to get bounds on the randomized error that do not depend on the distribution of the eigenvalues. We prove (see section 3) that for every k and p there are matrices for which the randomized error is very close to 1. This means that there are matrices for which the power method fails after k steps even for a random starting vector. In contrast to the problem of approximating the largest eigenvalue, this shows that the randomized error for the problem of approximating a largest eigenvector must depend on the distribution of eigenvalues. In particular, it must depend on the ratio between the two largest eigenvalues. So, the problem of approximating a largest eigenvector is harder than the problem of approximating the largest eigenvalue, and even a random start does not help to obtain distribution-free bounds.

We show that the rate of convergence of the power method depends on the ratio of the two largest eigenvalues, on their multiplicities, and on the particular norm $\mathcal{L}_p$. Let λ_1 be the largest eigenvalue with multiplicity r, and let λ_{r+1} be the second largest eigenvalue with multiplicity s. Then the randomized error after k steps is proportional to $(\lambda_{r+1}/\lambda_1)^k$ if $p < r$, to $k^{1/p}\ (\lambda_{r+1}/\lambda_1)^k$ if $p = r$, and to $(\lambda_{r+1}/\lambda_1)^{kr/p}$ if $p > r$. The multiplicative constants depend on p, r, and s. This means that the rate decreases with p, increases with multiplicity r, decreases with multiplicity s, and is at most linear in λ_{r+1}/λ_1. For $p = +\infty$, the power method has the randomized error equal to one for all k.

The results in this paper provide useful insight into the behavior of the power method for eigenvector approximation when the initial vector is randomly chosen. Our bounds can be useful for determining the computational cost for achieving a prescribed accuracy for eigenvector estimate. In fact, the sharpness of our upper and lower bounds allows one to derive an accurate estimate of the computational cost when the distribution of the eigenvalues is partially known. Another interesting result of the paper is that in some cases the randomized error has a rate of convergence lower than the well-known $(\lambda_{r+1}/\lambda_1)^k$ ratio achieved in the deterministic case. This is undoubtedly to be taken into account when one applies the power method with an initial starting vector.

We briefly comment on related work on approximate computation of eigenvectors. The idea of using random starting vectors for the power method can be found in Shub [8]. Shub applies the power method to the matrix e^{-A} and approximates an eigenvector of A which is *not* necessarily a largest eigenvector. Although, for this problem, the power method is globally convergent, the random start is used to improve efficiency. Shub shows, however, that even for $n = 2$ there are matrices for which this problem is very hard. In our paper we apply the power method to the matrix A, and we are only interested in approximating a largest eigenvector.

Kostlan [2] studies the randomized performance of the power method. In particular, in that paper he bounds the number of steps that allows the error to be lower than a fixed threshold ε. We discuss those bounds in section 4.3, comparing them with the bounds proposed in this paper.

Wright [10] and Kostlan [3] analyzed the problem of approximating a largest eigenvector by the power method in a different setting. They considered the average case setting over a class of matrices, whereas we consider the randomized setting.

In particular, they estimate the average time needed for computing a vector whose relative distance from the eigenspace of largest eigenvectors is less than ε. In our paper the matrix is fixed while the starting vector is chosen at random.

The paper is organized as follows. Section 2 contains the definition of the problem and some general results that are used in subsequent sections. In section 3 we analyze the behavior of the power method for worst case matrices. In section 4 we find upper and lower bounds on the randomized error. We show that these bounds are asymptotically optimal since, up to lower order terms, they match the asymptotic bounds presented in section 5. Numerical tests are presented in section 6. The tests show that the randomized error indeed depends on the distribution of the eigenvalues. We compare the test results with the theoretical lower and upper bounds. Section 7 contains the conclusions and final remarks.

2. Definition of the problem. Let A be an $n \times n$ symmetric positive definite matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_n > 0$ and corresponding orthonormal eigenvectors $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$. We will denote by $\mathcal{Z}$ the eigenspace corresponding to λ_1. We recall that the power method is defined as follows; see, e.g., [7]. Let $\mathbf{u}_0 = \mathbf{b}$ be any nonzero starting vector. Then, for every $k = 1, 2, \ldots$, we construct the following sequences of vectors:

$$\begin{cases} \mathbf{y}_k = A\mathbf{u}_{k-1}, \\ \mathbf{u}_k = \mathbf{y}_k / \|\mathbf{y}_k\|, \end{cases}$$

where $\|\cdot\|$ is the Euclidean vector norm.

Without loss of generality, we may assume that the starting vector $\mathbf{b}$ is normalized, so $\|\mathbf{b}\| = 1$. Observe that if we express $\mathbf{b}$ as a linear combination of the orthonormal eigenvectors,

$$\mathbf{b} = \sum_{i=1}^{n} b_i \mathbf{z}_i,$$

then $\mathbf{u}_k$ becomes

$$\mathbf{u}_k = \frac{\sum_{i=1}^{n} b_i \lambda_i^k \mathbf{z}_i}{\sqrt{\sum_{i=1}^{n} b_i^2 \lambda_i^{2k}}}. \tag{2.1}$$

Let r be the multiplicity of the largest eigenvalue λ_1. Without loss of generality, we assume that $1 \leq r < n$, since $r = n$ implies $A = \lambda_1 I$, and in this case any nonzero vector is an eigenvector corresponding to λ_1.

In order to estimate the error at the kth step, we consider the acute angle $\alpha_k(\mathbf{b})$ between the vector $\mathbf{u}_k$ and the eigenspace $\mathcal{Z}$. This angle is uniquely determined by the vector $\mathbf{u}_k$ and by its orthogonal projection on the subspace $\mathcal{Z}$. The sine of $\alpha_k(\mathbf{b})$ is the distance between the vector $\mathbf{u}_k$ and the subspace $\mathcal{Z}$. From (2.1) we have

$$\operatorname{dist}(\mathbf{u}_k, \mathcal{Z}) := \inf_{z \in \mathcal{Z}} \|\mathbf{u}_k - z\| = \sin(\alpha_k(\mathbf{b})) = \sqrt{\frac{\sum_{i=r+1}^{n} b_i^2 \lambda_i^{2k}}{\sum_{i=1}^{r} b_i^2 \lambda_1^{2k} + \sum_{i=r+1}^{n} b_i^2 \lambda_i^{2k}}}. \tag{2.2}$$

It is straightforward to see that, if the vector $\mathbf{b}$ has zero components in the directions of the eigenvectors belonging to λ_1 (i.e., $b_i = 0$ for $i = 1, 2 \ldots, r$), then $\alpha_k = \pi/2$ for any k. Otherwise, $\mathbf{u}_k$ converges to a vector of $\mathcal{Z}$ and the angle α_k goes to zero as k goes to infinity. The analysis of the power method for a fixed starting vector $\mathbf{b}$ may

be found in many books (see, for example, [7] and [9]), where in particular one finds that, if the method converges, the rate convergence is λ_{r+1}/λ_1.

As already mentioned, we study the randomized error of $\sin(\alpha_k(\cdot))$ in the $\mathcal{L}_p$ sense. Using (2.2) we have

$$\sin(\alpha_k(\mathbf{b})) = \sqrt{\frac{\sum_{i=r+1}^{n} b_i^2 x_i^{2k}}{\sum_{i=1}^{r} b_i^2 + \sum_{i=r+1}^{n} b_i^2 x_i^{2k}}}, \tag{2.3}$$

where

$$x_i = \lambda_i/\lambda_1 \quad \text{for } i = 1, 2, \ldots, n, \quad \text{and } 1 = x_1 = \cdots = x_r > x_{r+1} \cdots \geq x_n > 0. \tag{2.4}$$

Let us formalize the notion of $\mathcal{L}_p$ norm. Let μ be the uniform distribution over the unit sphere $S_n = \{\mathbf{b} : \|\mathbf{b}\| = 1\}$ such that $\mu(S_n) = 1$. Then the $\mathcal{L}_p$ norm of the function $\sin(\alpha_k(\cdot))$, defined as in (2.3), is given by

$$\|\sin(\alpha_k(\cdot))\|_p = \left[\int_{S_n} |\sin(\alpha_k(b))|^p \, \mu(db)\right]^{1/p}. \tag{2.5}$$

From Remark 7.2 of [4], we have

$$\int_{S_n} |\sin(\alpha_k(b))|^p \, \mu(db) = \frac{1}{c_n} \int_{B_n} |\sin(\alpha_k(b))|^p \, db, \tag{2.6}$$

where c_n is Lebesgue's measure of the unit ball $B_n = \{\mathbf{b} : \|\mathbf{b}\| \leq 1\}$; see (2.10) for the definition of c_n. Substituting (2.3) into (2.5) and using (2.6), we have

$$\|\sin(\alpha_k(\cdot))\|_p = \left[\frac{1}{c_n} \int_{B_n} \left(\frac{\sum_{i=r+1}^{n} b_i^2 x_i^{2k}}{\sum_{i=1}^{r} b_i^2 + \sum_{i=r+1}^{n} b_i^2 x_i^{2k}}\right)^{p/2} db\right]^{1/p}.$$

In the same way we define the norm of the space $\mathcal{L}_\infty$ to be

$$\begin{aligned} \|\sin(\alpha_k(\cdot))\|_\infty &= \sup_{\mathbf{b} \in S_n} |\sin(\alpha_k(b))| \\ &= \sup_{\|\mathbf{b}\|=1} \sqrt{\frac{\sum_{i=r+1}^{n} b_i^2 x_i^{2k}}{\sum_{i=1}^{r} b_i^2 + \sum_{i=r+1}^{n} b_i^2 x_i^{2k}}}. \end{aligned} \tag{2.7}$$

It is easy to see that the supremum in (2.7) is achieved by setting $\sum_{i=1}^{r} b_i^2 = 0$. From (2.7), we get

$$\|\sin(\alpha_k)\|_\infty = 1. \tag{2.8}$$

In the following we refer to $\sin(\alpha_k(\mathbf{b}))$ as the *error* of the power method after k steps for the starting vector $\mathbf{b}$. We denote $\|\sin(\alpha_k)\|_p$ by $e_k^{\text{ran}}(A, p)$, and we call it the *randomized error in the $\mathcal{L}_p$ sense* of the power algorithm after k steps. Hence, we have

$$e_k^{\text{ran}}(A, p) = \left[\frac{1}{c_n} \int_{B_n} \left(\frac{\sum_{i=r+1}^{n} b_i^2 x_i^{2k}}{\sum_{i=1}^{r} b_i^2 + \sum_{i=r+1}^{n} b_i^2 x_i^{2k}}\right)^{p/2} db\right]^{1/p}. \tag{2.9}$$

For $p = +\infty$, the power method fails to converge since its randomized error is 1 for all k; see (2.8). From now on we therefore assume that $p < +\infty$. As we shall see, the power method is then convergent: $e_k^{\mathrm{ran}}(A, p) \to 0$. The speed of convergence is, however, poor for large p.

In the paper we will denote by c_i the measure of the unit ball over $\mathbb{R}^i$. We have

$$c_i = \frac{\pi^{i/2}}{\Gamma\left(i/2+1\right)}; \tag{2.10}$$

see [1, eq. 8.310, 1] for the definition of the gamma function $\Gamma(x)$. We will also use the following relation between the beta and gamma functions:

$$B(i,j) = 2\int_0^1 t^{2i-1}(1-t^2)^{j-1}\,dt = \frac{\Gamma(i)\,\Gamma(j)}{\Gamma(i+j)}. \tag{2.11}$$

We will denote by $F(a,b;c;x)$ the hypergeometric function; see [1, eq. 9.10] for the definition and the properties of this function.

3. Worst case matrices. In [4], Kuczyński and Woźniakowski considered the power method for approximating the largest eigenvalue λ_1. They proved that the randomized error after k steps is bounded by a quantity that goes to zero as fast as $\ln(n)/k$ independently of the distribution of the eigenvalues. Our first goal is to analyze the possibility of obtaining distribution-free bounds for the problem of approximating a largest eigenvector. To this extent, we will deal with "worst case matrices."

Let us denote by $s(k,p)$ the supremum of the randomized error in the $\mathcal{L}_p$ sense over all positive definite matrices A, i.e.,

$$s(k,p) = \sup_{A=A^*>0} e_k^{\mathrm{ran}}(A,p).$$

Since the randomized error increases with x_i (see (2.4)), it is easy to show that the supremum is achieved by setting $x_i = 1$ for every $i \geq 2$ and for every p, $1 \leq p < \infty$. Then we get

$$\begin{aligned} s(k,p) &= \left[\int_{S_n}\left(\frac{\sum_{i=2}^n b_i^2}{b_1^2+\sum_{i=2}^n b_i^2}\right)^{p/2} db\right]^{1/p} \\ &= \left[\frac{1}{c_n}\int_{B_n}\left(1-\frac{b_1^2}{\sum_{i=1}^n b_i^2}\right)^{p/2} db\right]^{1/p}. \end{aligned} \tag{3.1}$$

Hence, $s(k,p)$ is independent of k and cannot go to zero. This shows that there are no distribution-free bounds. In fact, $s(k,p)$ are pretty close to 1. We first consider the case $p = 1$. Using (3.1) and the symmetry argument, we have

$$\begin{aligned} s(k,1) &= \frac{1}{c_n}\int_{B_n}\left(1-\frac{b_1^2}{\sum_{i=1}^n b_i^2}\right)^{1/2} db \\ &\geq \frac{1}{c_n}\int_{B_n}\left(1-\frac{b_1^2}{\sum_{i=1}^n b_i^2}\right) db = \left(1-\frac{1}{n}\right). \end{aligned} \tag{3.2}$$

We obtain estimates on $s(k,p)$ by the following proposition.

PROPOSITION 3.1. *For every k and p, $1 \le p < \infty$, we have*

$$\left(1 - \frac{1}{n}\right) \le s(k,p) \le 1.$$

Proof. The right-hand side inequality is trivial. Let us prove the left-hand side. For $p = 1$ it follows immediately by (3.2). For $p > 1$, applying Hölder's inequality to (3.1) gives us

$$\int_{S_n} \left(1 - \frac{b_1^2}{\sum_{i=1}^n b_i^2}\right)^{1/2} db \le \left[\int_{S_n} \left(1 - \frac{b_1^2}{\sum_{i=1}^n b_i^2}\right)^{p/2} db\right]^{1/p} \left[\int_{S_n} db\right]^{1/q},$$

where p and q are conjugate exponents; i.e., $1/p + 1/q = 1$. The proof is completed by observing that $\int_{S_n} db = 1$. □

Proposition 3.1 states that for every k there are matrices for which the randomized error is close to 1. These matrices have the largest eigenvalue of multiplicity 1, and the second largest eigenvalue has multiplicity $n-1$ and is pathologically close to λ_1. In this case, even if the starting vector is random, the sequence $\{\mathbf{u}_i\}$ for $i = 1, 2 \ldots, k$ does not approximate a largest eigenvector.

4. Nonasymptotic behavior. So far we have seen that if $\lambda_{r+1}/\lambda_1 \approx 1$ then the power method behaves badly even for a random starting vector. We now analyze the relationship between the ratio λ_{r+1}/λ_1 and the rate of convergence of the power method for approximating a largest eigenvector. We first show upper and lower bounds on the randomized error $e_k^{\mathrm{ran}}(A,p)$. These bounds depend on the distribution of the eigenvalues of the matrix A and on the particular norm used. In particular, we prove that the rate of convergence is slower when the multiplicity of λ_1 is smaller than the value of the parameter p of the norm. What seems interesting about these results is that they hold for a complete class of norms, and we are able to show how the speed of convergence of the power method depends on the norm.

4.1. Upper bounds. We now show how the rate of convergence depends on the multiplicity r of the largest eigenvalue and on the value of the parameter p of the norm. Theorem 4.1 shows that the rate of convergence depends on the relation between the parameters r and p. In particular, the speed of convergence increases with r and decreases with p.

THEOREM 4.1. *Let A be a symmetric positive definite matrix, and let r, $r < n$, denote the multiplicity of the largest eigenvalue λ_1 of A. Let*

$$\beta = \left[\frac{\Gamma(n/2)}{\Gamma(p/2)\,\Gamma((n-p)/2)} \left(2 + \frac{2}{n}\right)\right]^{1/p} x_{r+1}^k.$$

Then, for every p, $1 \le p < \infty$, and for every k we have

$$e_k^{\mathrm{ran}}(A,p) \le \begin{cases} x_{r+1}^k \left(\dfrac{\Gamma((r-p)/2)\,\Gamma((n+p-r)/2)}{\Gamma(r/2)\,\Gamma((n-r)/2)}\right)^{1/p} & \text{if } p < r, \\ x_{r+1}^k\,(2k)^{1/p} \left(\ln\left(\dfrac{1}{x_{r+1}}\right) \dfrac{\Gamma(n/2)}{\Gamma(p/2)\,\Gamma((n-p)/2)}\right)^{1/p} + \beta & \text{if } p = r, \\ x_{r+1}^{kr/p} \left(\dfrac{\Gamma((p-r)/2)\,\Gamma(n/2)}{\Gamma(p/2)\,\Gamma((n-r)/2)}\right)^{1/p} & \text{if } p > r. \end{cases}$$

Proof. We have

$$[e_k^{\mathrm{ran}}(A,p)]^p = \frac{1}{c_n}\int_{B_n}\left(\frac{\sum_{i=r+1}^n b_i^2 x_i^{2k}}{\sum_{i=1}^r b_i^2 + \sum_{i=r+1}^n b_i^2 x_i^{2k}}\right)^{p/2} db.$$

Observe that the integrand is an increasing function of $\sum_{i=r+1}^n b_i^2 x_i^{2k}$. The upper bound is then obtained by replacing x_i by x_{r+1} for $i > r+1$,

$$[e_k^{\mathrm{ran}}(A,p)]^p \le \frac{x_{r+1}^{kp}}{c_n}\int_{B_n}\left(\frac{\sum_{i=r+1}^n b_i^2}{\sum_{i=1}^r b_i^2 + x_{r+1}^{2k}\sum_{i=r+1}^n b_i^2}\right)^{p/2} db. \tag{4.1}$$

Let $a = x_{r+1}^k$, $\|b\|^2 = \sum_{i=1}^r b_i^2$, and let $t_i = b_i/(1-\|b\|^2)^{1/2}$ for $i = r+1,\dots,n$ with $\|t\|^2 = \sum_{i=r+1}^n t_i^2$. If we rewrite the last integral as an integral over the balls B_r and B_{n-r}, we get

$$[e_k^{\mathrm{ran}}(A,p)]^p \le \frac{a^p}{c_n}\int_{B_r}\int_{B_{n-r}} \frac{\|t\|^p\,(1-\|b\|^2)^{(n+p-r)/2}}{(\|b\|^2 + a^2\|t\|^2(1-\|b\|^2))^{p/2}}\,dt\,db.$$

Let $\gamma = r(n-r)c_r c_{n-r}/c_n$. We twice apply [1, eq. 4.642] to reduce the last integral to the two-dimensional integral, and we get

$$[e_k^{\mathrm{ran}}(A,p)]^p \le a^p\gamma\int_0^1\int_0^1 \frac{t^{n+p-r-1}\,b^{r-1}(1-b^2)^{(n+p-r)/2}}{(b^2 + a^2t^2(1-b^2))^{p/2}}\,db\,dt.$$

Since $b^2 + a^2t^2(1-b^2) \ge b^2$, we have

$$\begin{aligned}[e_k^{\mathrm{ran}}(A,p)]^p &\le a^p\gamma\int_0^1 t^{n+p-r-1}\,dt\int_0^1 \frac{b^{r-1}(1-b^2)^{(n+p-r-1)/2}}{b^p}\,db\\ &= a^p\frac{\gamma}{n+p-r}\int_0^1 b^{r-p-1}(1-b^2)^{(n+p-r)/2}\,db. \end{aligned}\tag{4.2}$$

Consider first the case $p < r$. From the definition of the beta function, (4.2) becomes

$$\begin{aligned}[e_k^{\mathrm{ran}}(A,p)]^p &\le a^p\frac{\gamma}{2(n+p-r)}B\left(\frac{r-p}{2},\frac{n+p-r}{2}+1\right)\\ &= a^p\frac{\Gamma\left((n+p-r)/2\right)\Gamma\left((r-p)/2\right)}{\Gamma\left((n-r)/2\right)\Gamma\left(r/2\right)}.\end{aligned}$$

This proves the case $p < r$.

Let us now consider the case $p = r$. The integral in (4.1) can be rewritten with respect to the ball B_{n-p} and the p-dimensional ball $B_p' = \{b : \sum_{i=1}^p b_i^2 \le 1 - \sum_{i=p+1}^n b_i^2\}$. We have

$$[e_k^{\mathrm{ran}}(A,p)]^p \le \frac{a^p}{c_n}\int_{B_{n-p}}\left(\sum_{i=p+1}^n b_i^2\right)^{p/2}\int_{B_p'}\frac{1}{\left(\sum_{i=1}^p b_i^2 + a^2\sum_{i=p+1}^n b_i^2\right)^{p/2}}\,db.$$

Let $\|b\|^2 = \sum_{i=p+1}^n b_i^2$. From [1, eq. 4.642], we get

$$[e_k^{\mathrm{ran}}(A,p)]^p \le a^p\frac{p\,c_p}{c_n}\int_{B_{n-p}}\|b\|^p\int_0^{\sqrt{1-\|b\|^2}}\frac{t^{p-1}}{(t^2 + a^2\|b\|^2)^{p/2}}\,dt\,db. \tag{4.3}$$

We have two cases, $p = r = 1$ and $p = r \geq 2$. If $p = 1$, (4.3) becomes

$$[e_k^{\mathrm{ran}}(A,1)] \leq a\,\frac{2}{c_n}\int_{B_{n-1}} \|b\| \int_0^{\sqrt{1-\|b\|^2}} \frac{1}{(t^2+a^2\|b\|^2)^{1/2}}\,dt\,db$$
$$= a\,\frac{2}{c_n}\int_{B_{n-1}} \|b\| \ln\left(\frac{\sqrt{1-\|b\|^2}+\sqrt{1-(1-a^2)\|b\|^2}}{a\,\|b\|}\right) db.$$

Using [1, eq. 4.642], and observing that $\sqrt{1-\|b\|^2} \leq \sqrt{1-(1-a^2)\|b\|^2}$, we get

$$[e_k^{\mathrm{ran}}(A,1)] \leq a\,\gamma \int_0^1 b^{n-1} \ln\left(\frac{2\sqrt{1-(1-a^2)b^2}}{ab}\right)$$
$$\leq a\,\frac{\gamma}{2n}\ln\left(\frac{1}{a^2}\right) + a\,\frac{\gamma}{n} + a\,\frac{\gamma}{n^2}, \tag{4.4}$$

where $\gamma = (n-1)2c_{n-1}/c_n$. Hence, from (4.4) we have

$$[e_k^{\mathrm{ran}}(A,1)] \leq a\,\frac{\Gamma(n/2)}{\Gamma(1/2)\,\Gamma((n-1)/2)}\,\ln\left(\frac{1}{a^2}\right) + a\,\frac{\Gamma(n/2)}{\Gamma(1/2)\,\Gamma((n-1)/2)}\left(2+\frac{2}{n}\right).$$

This proves the case $p = r = 1$.

Let us consider the case $p \geq 2$. Notice that

$$(t^2+a^2\|b\|^2)^{p/2} \geq t^p + \frac{p}{2}t^{2(p/2-1)}a^2\|b\|^2.$$

Then we can bound the denominator of the integrand of (4.3) with the first two terms of this expansion. We have

$$[e_k^{\mathrm{ran}}(A,p)]^p \leq a^p\frac{p\,c_p}{c_n}\int_{B_{n-p}} \|b\|^p \int_0^{\sqrt{1-\|b\|^2}} \frac{t^{p-1}}{t^p + p/2\,t^{p-2}a^2\|b\|^2}\,dt\,db$$
$$= a^p\frac{p\,c_p}{c_n}\int_{B_{n-p}} \|b\|^p \int_0^{\sqrt{1-\|b\|^2}} \frac{t}{t^2 + p/2\,a^2\|b\|^2}\,dt\,db.$$

Solving the last integral, and using again [1, eq. 4.642] to reduce the first integral to a one-dimensional integral, we obtain

$$[e_k^{\mathrm{ran}}(A,p)]^p \leq a^p\gamma\int_0^1 b^{n-1}\,\frac{1}{2}\ln\left(\frac{1-(1-p/2a^2)b^2}{p/2\,a^2b^2}\right) db$$
$$= a^p\frac{\gamma}{2}\int_0^1 b^{n-1}\ln\left(\frac{1}{p/2\,a^2b^2}\right) db + a^p\frac{\gamma}{2}\int_0^1 b^{n-1}\ln\left(1-\left(1-\frac{p}{2}a^2\right)b^2\right) db$$
$$= a^p\frac{\gamma}{2n}\ln\left(\frac{2}{pa^2}\right) + a^p\frac{\gamma}{n^2} + a^p\frac{\gamma}{2}\int_0^1 b^{n-1}\ln\left(1-\left(1-\frac{p}{2}a^2\right)b^2\right) db, \tag{4.5}$$

where $\gamma = p(n-p)c_pc_{n-p}/c_n$. Let us consider the argument of the logarithm in the integral of (4.5). Observe that if $a^2 \leq 2/p$, then $\ln(1-(1-p/2\,a^2)b^2) \leq 0$. Hence, in this case, we can bound (4.5) by

$$[e_k^{\mathrm{ran}}(A,p)]^p \leq a^p\frac{\gamma}{2n}\ln\left(\frac{2}{pa^2}\right) + a^p\frac{\gamma}{n^2}$$

$$= a^p \frac{\gamma}{2n} \ln\left(\frac{1}{a^2}\right) + a^p \frac{\gamma}{2n}\left(\ln\left(\frac{2}{p}\right) + \frac{2}{n}\right)$$

$$\leq a^p \frac{\Gamma(n/2)}{\Gamma(p/2)\,\Gamma((n-p)/2)} \ln\left(\frac{1}{a^2}\right) + a^p \frac{\Gamma(n/2)}{\Gamma(p/2)\,\Gamma((n-p)/2)} \frac{2}{n}. \tag{4.6}$$

Otherwise, if $a^2 > 2/p$, then $\ln(1-(1-p/2\,a^2)b^2) \leq \ln(p/2)$. In this case we have

$$[e_k^{\mathrm{ran}}(A,p)]^p \leq a^p \frac{\gamma}{2n} \ln\left(\frac{2}{pa^2}\right) + a^p \frac{\gamma}{n^2} + a^p \frac{\gamma}{2} \int_0^1 b^{n-1} \ln\left(\frac{p}{2}\right) db$$

$$= a^p \frac{\gamma}{2n} \ln\left(\frac{1}{a^2}\right) + a^p \frac{\gamma}{n^2}$$

$$= a^p \frac{\Gamma(n/2)}{\Gamma(p/2)\,\Gamma((n-p)/2)} \ln\left(\frac{1}{a^2}\right) + a^p \frac{\Gamma(n/2)}{\Gamma(p/2)\,\Gamma((n-p)/2)} \frac{2}{n}. \tag{4.7}$$

Observing that $2/n < (2+2/n)$ and using (4.6) and (4.7), we have

$$[e_k^{\mathrm{ran}}(A,p)]^p \leq a^p \frac{\Gamma(n/2)}{\Gamma(p/2)\,\Gamma((n-p)/2)} \ln\left(\frac{1}{a^2}\right) + a^p \frac{\Gamma(n/2)}{\Gamma(p/2)\,\Gamma((n-p)/2)} \left(2+\frac{2}{n}\right).$$

This proves the case $p = r$.

Finally, assume that $p > r$. From (4.1), repeating the same reasoning that led to (4.3), we have

$$[e_k^{\mathrm{ran}}(A,p)] \leq a^p \frac{rc_r}{c_n} \int_{B_{n-r}} \|b\|^p \int_0^{\sqrt{1-\|b\|^2}} \frac{t^{r-1}}{(t^2+a^2\|b\|^2)^{p/2}}\, dt\, db$$

$$= \frac{rc_r}{c_n} \int_{B_{n-r}} \int_0^{\sqrt{1-\|b\|^2}} \frac{t^{r-1}}{(t^2/(a^2\|b\|^2)+1)^{p/2}}\, dt\, db.$$

Changing variables by setting $z = t/(a\|b\|)$, we get

$$[e_k^{\mathrm{ran}}(A,p)]^p \leq a^r \frac{rc_r}{c_n} \int_{B_{n-r}} \|b\|^r \int_0^d \frac{z^{r-1}}{(z^2+1)^{p/2}}\, dz\, db,$$

where $d = \sqrt{1-\|b\|^2}/(a\|b\|)$. Now set $y = z^2$. From the last equation we have

$$[e_k^{\mathrm{ran}}(A,p)]^p \leq a^r \frac{rc_r}{2c_n} \int_{B_{n-r}} \|b\|^r \int_0^{d^2} \frac{y^{r/2-1}}{(y+1)^{p/2}}\, dy\, db.$$

We notice that d goes to infinity when a goes to zero. Then we have

$$\int_0^{d^2} \frac{y^{r/2-1}}{(y+1)^{p/2}}\, dy \leq \int_0^{+\infty} \frac{y^{r/2-1}}{(y+1)^{p/2}}\, dy = B\left(\frac{r}{2}, \frac{p-r}{2}\right),$$

due to formula [1, eq. 3.194]. We apply [1, eq. 4.642] to reduce the integral over B_{n-r} to a one-dimensional integral, and we get

$$[e_k^{\mathrm{ran}}(A,p)]^p \leq a^r \frac{r(n-r)c_r c_{n-r}}{c_n} \int_0^1 b^{n-1} B\left(\frac{r}{2}, \frac{p-r}{2}\right) = a^r \frac{\Gamma(n/2)\,\Gamma((p-r)/2)}{\Gamma((n-r)/2)\,\Gamma(p/2)}.$$

This concludes the proof. ∎

Note that, when $p = r$, the bound is composed of two terms. The first term depends on k through $x_{r+1}^k k^{1/p}$; the second term depends on k through x_{r+1}^k. We remark that for large k the influence of the second term is negligible. Nevertheless, numerical tests show that this term can affect the bound when the value of x_{r+1} is close to 1.

4.2. Lower bounds. In this section we find lower bounds on the randomized error $e_k^{\rm ran}(A,p)$. As in section 4.1, we show that these lower bounds depend on the multiplicity of the largest eigenvalue and on the value of the parameter p of the norm. Upper and lower bounds show the same dependence on the ratio between the two largest eigenvalues and on the relation between p and r.

Below we define some constants that are used in Theorem 4.2.

$$\gamma = \left(\frac{\Gamma\left((r-p)/2\right)\Gamma\left(p+1/2\right)\Gamma\left((r+1)/2\right)}{2\Gamma\left(r/2\right)\Gamma\left(1/2\right)\Gamma\left((r+p+1)/2\right)} F\Big(\frac{r+1}{2}, \frac{r-p}{2}; \frac{r+p+1}{2}; 1-x_{r+1}^{2k} \Big) \right)^{1/p}$$

if $p<2$, and

$$\gamma = \left(\frac{p\Gamma\left((r-p)/2\right)\Gamma\left((p+3)/2\right)}{2(r+1)\Gamma\left(r/2\right)\Gamma\left(1/2\right)} F\Big(\frac{r+1}{2}, \frac{r-p}{2}; \frac{r+3}{2}; 1-x_{r+1}^{2k} \Big) \right)^{1/p}$$

if $p \geq 2$. Moreover,

$$\gamma' = \left(\frac{\Gamma\left((p+1)/2\right)}{\Gamma\left(p/2\right)\Gamma\left(1/2\right)} \left(\log\left(\frac{p^2+3p}{4}\right) - \frac{2}{p+1} + \frac{2p-4}{p+3} F\Big(1, \frac{p+1}{2}; \frac{p+5}{2}; 1-x_{r+1}^{2k}\Big)\right)\right)^{1/p},$$

$$\gamma'' = \left(\frac{r\,\Gamma\left((r+1)/2+1\right)\Gamma\left((p-r)/2\right)}{4p\,\Gamma\left(1/2\right)\Gamma\left(p/2+1\right)} F\left(\frac{r}{2}+1, 1; \frac{p}{2}+1; 1-x_{r+1}^{2k}\right)\right)^{1/p}.$$

THEOREM 4.2. *Let A be a symmetric positive definite matrix, and let r, $r<n$, denote the multiplicity of the largest eigenvalue λ_1 of A. Then, for every p, $1 \leq p < \infty$, and for every k we have*

$$e_k^{\rm ran}(A,p) \geq \begin{cases} x_{r+1}^k \left(\dfrac{\Gamma\left((r-p)/2\right)\Gamma\left((p+1)/2\right)}{\Gamma\left(r/2\right)\Gamma\left(1/2\right)} \right)^{1/p} - \gamma\, x_{r+1}^{kr/p} & \text{if } p<r, \\[2ex] x_{r+1}^k (2k)^{1/r} \left(\ln\left(\dfrac{1}{x_{r+1}}\right) \dfrac{\Gamma\left((p+1)/2\right)}{\Gamma\left(p/2\right)\Gamma\left(1/2\right)} \right)^{1/p} - \gamma' x_{r+1}^k & \text{if } p=r, \\[2ex] x_{r+1}^{kr/p} \left(\dfrac{\Gamma\left((p-r)/2\right)\Gamma\left((r+1)/2\right)}{\Gamma\left(p/2\right)\Gamma\left(1/2\right)} \right)^{1/p} - \gamma'' x_{r+1}^{k(r+2)/p} & \text{if } p>r. \end{cases}$$

Proof. We have

$$[e_k^{\rm ran}(A,p)]^p = \frac{1}{c_n} \int_{B_n} \left(\frac{\sum_{i=r+1}^n b_i^2 x_i^{2k}}{\sum_{i=1}^r b_i^2 + \sum_{i=r+1}^n b_i^2 x_i^{2k}} \right)^{p/2} db.$$

Notice that the integrand is an increasing function of $\sum_{i=r+1}^n b_i^2 x_i^{2k}$. Hence, the lower bound is obtained by replacing x_i by 0 for $i > r+1$,

$$[e_k^{\rm ran}(A,p)]^p \geq \frac{x_{r+1}^{kp}}{c_n} \int_{B_n} \frac{b_{r+1}^p}{\left(\sum_{i=1}^r b_i^2 + x_{r+1}^{2k} b_{r+1}^2\right)^{p/2}} db.$$

Let $a = x_{r+1}^k$. Writing the last integral as an integral over the ball B_{n-r} and the r-dimensional ball of radius $q = \sqrt{1-\sum_{i=r+1}^n b_i^2}$, and applying [1, eq. 4.642], we get

$$[e_k^{\rm ran}(A,p)]^p \geq \frac{r c_r}{c_n} \int_{B_{n-r}} \int_0^q t^{r-1} \left(\frac{a^2\, b_{r+1}^2}{t^2 + a^2\, b_{r+1}^2} \right)^{p/2} dt\, db. \tag{4.8}$$

Let us denote $a^2\, b_{r+1}^2$ by α and consider the integral

$$f(\alpha) = \int_0^q t^{r-1} \left(\frac{\alpha}{t^2+\alpha} \right)^{p/2} dt. \tag{4.9}$$

We have three cases depending on the relation between p and r.

Consider first the case $p < r$. It is convenient to split $f(\alpha)$ as follows:

$$f(\alpha) = \alpha^{p/2} \left(\int_0^q t^{r-p-1}\, dt - \int_0^q g(t)\, dt \right), \tag{4.10}$$

where

$$g(t) = t^{r-1} \left(\frac{1}{t^p} - \left(\frac{1}{t^2+\alpha} \right)^{p/2} \right).$$

We can conveniently rewrite $g(t)$ as

$$g(t) = t^{r-p-1} \left(1 - \left(\frac{t^2}{t^2+\alpha} \right)^{p/2} \right).$$

Setting $y = t^2/\alpha$, we have

$$\int_0^q g(t)\, dt = \frac{\alpha^{(r-p)/2}}{2} \int_0^{q^2/\alpha} y^{(r-p)/2-1} \frac{(y+1)^{p/2} - y^{p/2}}{(y+1)^{p/2}}\, dy. \tag{4.11}$$

We consider two cases: $p < 2$ and $p \geq 2$. Let us start with $p < 2$. Notice that $(y+1)^{p/2} - y^{p/2} \leq 1$. Then from (4.11) we get

$$\begin{aligned} \int_0^q g(t)\, dt &\leq \frac{\alpha^{(r-p)/2}}{2} \int_0^{q^2/\alpha} \frac{y^{(r-p)/2-1}}{(y+1)^{p/2}}\, dy \\ &= \frac{q^{r-p}}{r-p} F\left(\frac{p}{2}, \frac{r-p}{2}; \frac{r-p}{2}+1; -\frac{q^2}{\alpha} \right), \end{aligned}$$

due to [1, eq. 3.194] (also see [1] for the definition and the properties of the hypergeometric function $F(a, b; c; x)$). Substituting it into (4.10) and solving the first integral, we have

$$f(\alpha) \geq \alpha^{p/2} \frac{q^{r-p}}{r-p} - \alpha^{p/2} \frac{q^{r-p}}{r-p} F\left(\frac{p}{2}, \frac{r-p}{2}; \frac{r-p}{2}+1; -\frac{q^2}{\alpha} \right).$$

Hence, (4.8) becomes

$$\begin{aligned} [e_k^{\mathrm{ran}}(A,p)]^p &\geq \frac{r c_r}{(r-p)c_n} \int_{B_{n-r}} a^p b_{r+1}^p \left(1 - \sum_{i=r+1}^n b_i^2 \right)^{(r-p)/2} db \\ &- \frac{a^p r c_r}{(r-p)c_n} \int_{B_{n-r}} b_{r+1}^p \left(1 - \sum_{i=r+1}^n b_i^2 \right)^{(r-p)/2} F\left(\frac{p}{2}, \frac{r-p}{2}; \frac{r-p+2}{2}; \frac{\sum_{i=r+1}^n b_i^2 - 1}{a^2 b_{r+1}^2} \right) db. \end{aligned}$$

Using [1, eq. 4.642], we get

$$[e_k^{\mathrm{ran}}(A,p)]^p \geq a^p \frac{\Gamma\left((p+1)/2\right)\Gamma\left((r-p)/2\right)}{\Gamma\left(1/2\right)\Gamma\left(r/2\right)} \tag{4.12}$$
$$- a^p \frac{(r+1)\Gamma\left((r+1)/2\right)}{(r-p)\Gamma\left(r/2\right)\Gamma\left(1/2\right)} \int_0^1 t^p (1-t^2)^{(r-p)/2} F\left(\frac{p}{2}, \frac{r-p}{2}; \frac{r-p+2}{2}; \frac{t^2-1}{a^2t^2}\right) dt.$$

After setting $y = (1-t^2)/(a^2t^2)$, we can rewrite the integral in (4.12) as

$$\int_0^1 t^p (1-t^2)^{(r-p)/2} F\left(\frac{p}{2}, \frac{r-p}{2}; \frac{r-p}{2}+1; -\frac{1-t^2}{a^2t^2}\right) dt$$
$$= \frac{a^{-p-1}}{2} \int_0^\infty \frac{y^{(r-p)/2}}{(y+1/a^2)^{(r+3)/2}} F\left(\frac{p}{2}, \frac{r-p}{2}; \frac{r-p}{2}+1; -y\right) dy.$$

From the last equation and using [1, eq. 7.512, 10], we have

$$\int_0^1 t^p (1-t^2)^{(r-p)/2} F\left(\frac{p}{2}, \frac{r-p}{2}; \frac{r-p}{2}+1; \frac{t^2-1}{a^2t^2}\right) dt \tag{4.13}$$

$$= \frac{2a^{p+1}\Gamma\left((r-p)/2+1\right)\Gamma\left(p+1/2\right)\Gamma\left((r+1)/2\right)}{\Gamma\left((r+3)/2\right)\Gamma\left((p+r+1)/2\right)} F\left(p+\frac{1}{2}, \frac{r+1}{2}; \frac{p+r+1}{2}; 1-\frac{1}{a^2}\right).$$

Applying the transformation formula to the hypergeometric function (see [1, eq. 9.131, 1]), we have

$$F\left(p+\frac{1}{2}, \frac{r+1}{2}; \frac{p+r+1}{2}; 1-\frac{1}{a^2}\right) = a^{r+1} F\left(\frac{r+1}{2}, \frac{r-p}{2}; \frac{p+r+1}{2}; 1-a^2\right).$$

Substituting it into (4.13) and then into (4.12), we get

$$[e_k^{\mathrm{ran}}(A,p)]^p \geq a^p \frac{\Gamma\left((p+1)/2\right)\Gamma\left((r-p)/2\right)}{\Gamma\left(1/2\right)\Gamma\left(r/2\right)} - a^r\gamma,$$

where

$$\gamma = \frac{\Gamma\left((r-p)/2\right)\Gamma\left(p+1/2\right)\Gamma\left((r+1)/2\right)}{2\Gamma\left(r/2\right)\Gamma\left(1/2\right)\Gamma\left((r+p+1)/2\right)} F\left(\frac{r+1}{2}, \frac{r-p}{2}; \frac{r+p+1}{2}, 1-a^2\right).$$

This concludes the proof of the case $p < 2$.

Let $p \geq 2$. Observe that, from Lagrange's theorem, there exists a value ξ, $y \leq \xi \leq y+1$, such that $(y+1)^{p/2} - y^{p/2} = p/2\, \xi^{p/2-1}$. Since $\xi^{p/2-1} \leq (y+1)^{p/2-1}$, we obtain the bound

$$\int_0^q g(t)\, dt \leq \frac{\alpha^{(r-p)/2}\, p}{4} \int_0^{q^2/\alpha} \frac{y^{(r-p)/2-1}}{y+1}\, dy$$
$$= q^{r-p} \frac{p}{2(r-p)} F\left(1, \frac{r-p}{2}; \frac{r-p}{2}+1; -\frac{q^2}{\alpha}\right), \tag{4.14}$$

which follows from [1, eq. 3.194, 1]. Proceeding exactly as before, we get

$$f(\alpha) \geq \alpha^{p/2} \frac{q^{r-p}}{r-p} - \alpha^{p/2} \frac{q^{r-p}p}{2(r-p)} F\left(1, \frac{r-p}{2}; \frac{r-p}{2}+1; -\frac{q^2}{\alpha}\right).$$

Using this bound in (4.8), we get

$$[e_k^{\mathrm{ran}}(A,p)]^p \geq \frac{rc_r}{(r-p)c_n}\int_{B_{n-r}} a^p b_{r+1}^p \left(1-\sum_{i=r+1}^{n} b_i^2\right)^{(r-p)/2} db \tag{4.15}$$

$$-\frac{rpc_r}{2(r-p)c_n}\int_{B_{n-r}} a^p b_{r+1}^p \left(1-\sum_{i=r+1}^{n} b_i^2\right)^{(r-p)/2} F\left(1,\frac{r-p}{2};\frac{r-p+2}{2};\frac{\sum_{i=r+1}^{n} b_i^2-1}{a^2 b_{r+1}^2}\right) db.$$

Solving the second integral in (4.15) as before and applying the transformation formula [1, eq. 9.131] to the hypergeometric function, we have

$$[e_k^{\mathrm{ran}}(A,p)]^p \geq a^p \frac{\Gamma\left((p+1)/2\right)\Gamma\left((r-p)/2\right)}{\Gamma\left(1/2\right)\Gamma\left(r/2\right)} - a^r\gamma,$$

where

$$\gamma = \frac{p\Gamma\left((r-p)/2\right)\Gamma\left((p+3)/2\right)}{2(r+1)\Gamma\left(r/2\right)\Gamma\left(1/2\right)} F\left(\frac{r+1}{2},\frac{r-p}{2};\frac{r+1}{2}+1,1-a^2\right).$$

This concludes the proof for $p<r$.

Let $p=r$. The integral denoted by $f(\alpha)$ in (4.9) becomes

$$f(\alpha) = \alpha^{p/2}\int_0^q t^{p-1}\left(\frac{1}{t^2+\alpha}\right)^{p/2} dt \tag{4.16}$$

and can be rewritten as

$$f(\alpha) = \alpha^{p/2}\left(\int_0^q \frac{t}{t^2+p/2\,\alpha}\,dt - \int_0^q g(t)\,dt\right), \tag{4.17}$$

where

$$g(t) = \frac{t}{t^2+p/2\,\alpha} - \frac{t^{p-1}}{(t^2+\alpha)^{p/2}}.$$

Since $p=r$, we have that p is an integer between 1 and n. We analyze separately the cases $p=1$ and $p\geq 2$. If $p=1$, then $g(t)\leq 0$ and

$$\begin{aligned} f(\alpha) &\geq \alpha^{1/2}\left(\int_0^q \frac{t}{t^2+1/2\,\alpha}\,dt\right) \\ &= \frac{\alpha^{1/2}}{2}\ln\left(\frac{q^2+1/2\,\alpha}{1/2\,\alpha}\right). \end{aligned}$$

From (4.8) and since $q=\sqrt{1-\sum_{i=2}^{n} b_i^2}$, we get

$$[e_k^{\mathrm{ran}}(A,1)]^1 \geq \frac{1}{c_n}\int_{B_{n-1}} \alpha^{1/2}\ln\left(\frac{1-\sum_{i=2}^{n} b_i^2+1/2\,\alpha}{1/2\,\alpha}\right) db.$$

Let $\|b\| = \sum_{i=3}^{n} b_i^2$ and $t = b_2/(1-\|b\|^2)^{1/2}$. Since $\alpha = a^2\, b_2^2$, using [1, eq. 4.642], we have

$$\begin{aligned}
[e_k^{\mathrm{ran}}(A,1)] &\geq a\,\frac{2}{c_n}\int_{B_{n-2}}\int_0^1 t\,(1-\|b\|^2)\ln\left(\frac{1-(1-1/2\,a^2)t^2}{1/2\,a^2t^2}\right)\,dt\,db\\
&= a\,\frac{(n-2)c_{n-2}}{c_n}B\left(\frac{n}{2}-1,2\right)\int_0^1 t\,\ln\left(\frac{1-(1-1/2\,a^2)t^2}{1/2a^2t^2}\right)\,dt\\
&= a\,\frac{(n-2)c_{n-2}}{c_n}B\left(\frac{n}{2}-1,2\right)\frac{1}{2(1-1/2\,a^2)}\ln\left(\frac{2}{a^2}\right)\\
&\geq a\frac{(n-2)c_{n-2}}{2c_n}B\left(\frac{n}{2}-1,2\right)\ln\left(\frac{1}{a^2}\right)+a\frac{(n-2)c_{n-2}}{2c_n}B\left(\frac{n}{2}-1,2\right)\ln(2),
\end{aligned}$$

from which we have

$$e_k^{\mathrm{ran}}(A,1) \geq \frac{a}{\pi}\ln\left(\frac{1}{a^2}\right)+\frac{a}{\pi}\ln(2).$$

This provides the proof for $p = r = 1$.

Now let $p \geq 2$. We notice that $t^2 + p/2\,\alpha \geq t^2 + \alpha$. Then

$$g(t) \leq \frac{t(t^2+\alpha)^{p/2-1} - t^{p-1}}{(t^2+\alpha)^{p/2}}.$$

Setting $y = t^2/\alpha$, we have

$$\begin{aligned}
\int_0^q g(t)\,dt &\leq \int_0^{q^2/\alpha}\frac{(y+1)^{p/2-1}-y^{p/2-1}}{2(y+1)^{p/2}}\,dy\\
&\leq \frac{1}{2}\left(\frac{p}{2}-1\right)\int_0^{q^2/\alpha}\frac{1}{(y+1)^2}\,dy\\
&= \frac{1}{2}\left(\frac{p}{2}-1\right)\frac{q^2}{\alpha+q^2}.
\end{aligned}$$

We substitute this inequality into (4.17). We have

$$\begin{aligned}
f(\alpha) &\geq \alpha^{p/2}\left(\int_0^q \frac{t}{t^2+p/2\,\alpha}\,dt-\frac{1}{2}\left(\frac{p}{2}-1\right)\right)\\
&= \frac{\alpha^{p/2}}{2}\ln\left(\frac{q^2+p/2\,\alpha}{p/2\,\alpha}\right)-\frac{\alpha^{p/2}}{2}\left(\frac{p}{2}-1\right)\frac{q^2}{\alpha+q^2}.
\end{aligned}$$

Since $q = \sqrt{1-\sum_{i=p+1}^{n} b_i^2}$ and $p = r$, we obtain the lower bound

$$\begin{aligned}
[e_k^{\mathrm{ran}}(A,p)]^p \geq{}& \frac{pc_p}{2c_n}\int_{B_{n-p}}\alpha^{p/2}\ln\left(\frac{1-\sum_{i=p+1}^{n} b_i^2+p/2\,\alpha}{p/2\,\alpha}\right)\,db\\
&-\frac{pc_p}{2c_n}\left(\frac{p}{2}-1\right)\int_{B_{n-p}}\alpha^{p/2}\frac{1-\sum_{i=r}^{n} b_i^2}{\alpha+1-\sum_{i=r}^{n} b_i^2}\,db.
\end{aligned}$$

Let $\|b\|^2 = \sum_{i=p+2}^{n} b_i^2$ and $t = b_{p+1}/(1-\|b\|^2)^{1/2}$. Then from the definition of α and

using [1, eq. 4.642], we have

$$[e_k^{\mathrm{ran}}(A,p)]^p \geq a^p \frac{pc_p}{c_n} \int_{B_{n-p-1}} \int_0^1 t^p (1-\|b\|^2)^{(p+1)/2} \ln\left(\frac{1-(1-p/2\,a^2)t^2}{p/2\,a^2t^2}\right) dt\,db$$
$$- a^p \frac{pc_p}{c_n}\left(\frac{p}{2}-1\right) \int_{B_{n-p-1}} \int_0^1 t^p (1-\|b\|^2)^{(p+1)/2} \frac{(1-t^2)}{1-(1-a^2)t^2}\,dt\,db.$$

Again using [1, eq. 4.642], we get

$$[e_k^{\mathrm{ran}}(A,p)]^p \geq a^p \gamma B\left(\frac{n-p-1}{2}, \frac{p+1}{2}+1\right) \int_0^1 t^p \ln\left(\frac{1-(1-p/2\,a^2)t^2}{p/2\,a^2t^2}\right) dt$$
$$- a^p \gamma \frac{p-2}{2} B\left(\frac{n-p-1}{2}, \frac{p+1}{2}+1\right) \int_0^1 \frac{t^p(1-t^2)}{1-(1-a^2)t^2}\,dt, \tag{4.18}$$

where $\gamma = p(n-p-1)c_pc_{n-p-1}/(2c_n)$. Observe that

$$\int_0^1 t^p \ln\left(\frac{1-(1-p/2\,a^2)t^2}{p/2\,a^2t^2}\right) dt$$
$$= \frac{1}{p+1}\ln\left(\frac{2}{pa^2}\right) + \frac{2}{(p+1)^2} + \int_0^1 t^p \ln\left(1-\left(1-\frac{p}{2}a^2\right)t^2\right) dt. \tag{4.19}$$

Notice that if $a^2 > 2/p$ then $\ln(1-(1-p/2\,a^2)t^2) \geq \ln(1) = 0$. Hence, from (4.19) and using [1, eq. 3.197, 3] to solve the integral in (4.18), we have

$$[e_k^{\mathrm{ran}}(A,p)]^p \geq a^p \frac{\gamma'}{p+1}\ln\left(\frac{2}{pa^2}\right)$$
$$- a^p\gamma' \frac{2(p-2)}{(p+3)} F\left(1, \frac{p+1}{2}; \frac{p+5}{2}; 1-a^2\right) + a^p\gamma' \frac{2}{p+1},$$

where $\gamma' = \gamma B\left((n-p-1)/2, (p+1)/2+1\right)$. Using (2.10), we can express c_i in terms of the gamma function, and we get

$$[e_k^{\mathrm{ran}}(A,p)]^p \geq a^p \frac{\Gamma\left((p+1)/2\right)}{\Gamma\left(p/2\right)\Gamma\left(1/2\right)} \ln\left(\frac{1}{a^2}\right)$$
$$- a^p \frac{\Gamma\left((p+1)/2\right)}{\Gamma\left(p/2\right)\Gamma\left(1/2\right)} \left[\frac{2(p-2)}{p+3} F\left(1, \frac{p+1}{2}; \frac{p+5}{2}; 1-a^2\right) - \ln\left(\frac{2}{p}\right) - \frac{2}{p+1}\right]. \tag{4.20}$$

Otherwise, when $a^2 \leq 2/p$, we use the fact that

$$\ln(1-ct^2) = -\sum_{i=1}^{\infty} \frac{(ct^2)^i}{i},$$

where c is a constant such that $-1 \leq ct^2 < 1$. Setting $c = (1-p/2a^2)$, and using the previous relation, the integral in (4.19) becomes

$$\int_0^1 t^p \ln\left(1-\left(1-\frac{p}{2}a^2\right)t^2\right) dt = -\sum_{i=1}^{\infty} \frac{(1-p/2\,a^2)^i}{i(2i+p+1)} \geq -\frac{1}{p+1}\ln\left(\frac{p+3}{2}\right).$$

In this case, from (4.19) we have

$$\int_0^1 t^p \ln\left(\frac{1-(1-p/2a^2)t^2}{p/2a^2t^2}\right)\,dt \geq \frac{1}{p+1}\ln\left(\frac{2}{pa^2}\right) - \frac{1}{p+1}\ln\left(\frac{p+3}{2}\right) + \frac{2}{(p+1)^2},$$

and then

$$[e_k^{\mathrm{ran}}(A,p)]^p \geq a^p \frac{\Gamma\left((p+1)/2\right)}{\Gamma\left(p/2\right)\Gamma\left(1/2\right)}\ln\left(\frac{1}{a^2}\right)$$
$$- a^p \frac{\Gamma\left((p+1)/2\right)}{\Gamma\left(p/2\right)\Gamma\left(1/2\right)}\left[\frac{2p-4}{p+3}F\left(1,\frac{p+1}{2};\frac{p+5}{2};1-a^2\right) + \ln\left(\frac{p(p+3)}{4}\right) - \frac{2}{p+1}\right],$$

which concludes the proof for $p = r$.

The last case is $p > r$. Setting $y = t^2/\alpha$, the integral $f(\alpha)$ defined by (4.9) becomes

$$f(\alpha) = \frac{\alpha^{r/2}}{2}\int_0^{q^2/\alpha} \frac{y^{r/2-1}}{(y+1)^{p/2}}\,dy.$$

It can be rewritten as

$$f(\alpha) = \frac{\alpha^{r/2}}{2}\left[\int_0^{\infty} \frac{y^{r/2-1}}{(y+1)^{p/2}}\,dy - \int_{q^2/\alpha}^{\infty} \frac{y^{r/2-1}}{(y+1)^{p/2}}\,dy\right]. \tag{4.21}$$

The first integral of the right-hand side of (4.21) can be solved using [1, eq. 3.194, 3] and is equal to $B(r/2,(p-r)/2)$. The second integral of (4.21) can be solved using [1, eq. 3.194, 2] and is equal to

$$\left(\frac{\alpha}{q^2}\right)^{(p-r)/2}\frac{2}{p-r}F\left(\frac{p}{2},\frac{p-r}{2};\frac{p-r}{2}+1;-\frac{\alpha}{q^2}\right).$$

Hence, (4.21) becomes

$$f(\alpha) = \frac{\alpha^{r/2}}{2}B\left(\frac{r}{2},\frac{p-r}{2}\right) - \frac{\alpha^{p/2}}{(p-r)\,q^{p-r}}F\left(\frac{p}{2},\frac{p-r}{2};\frac{p-r}{2}+1;-\frac{\alpha}{q^2}\right). \tag{4.22}$$

By substituting (4.22) into (4.8), and from the definition of α and q we have

$$[e_k^{\mathrm{ran}}(A,p)]^p \geq a^r \frac{rc_r}{2c_n}B\left(\frac{r}{2},\frac{p-r}{2}\right)\int_{B_{n-r}} b_{r+1}^r\,db$$
$$- \frac{a^p r c_r}{(p-r)c_n}\int_{B_{n-r}} \frac{b_{r+1}^p}{\left(1-\sum_{i=r+1}^n b_i^2\right)^{(p-r)/2}}F\left(\frac{p}{2},\frac{p-r}{2};\frac{p-r+2}{2};\frac{a^2 b_{r+1}^2}{\sum_{i=r+1}^n b_i^2 - 1}\right)db.$$

Using again the technique of reducing integrals to one-dimensional integrals, we get

$$[e_k^{\mathrm{ran}}(A,p)]^p \geq a^r\frac{\gamma}{2}B\left(\frac{r}{2},\frac{p-r}{2}\right)\,B\left(\frac{r+1}{2},\frac{n-r-1}{2}+1\right)$$
$$- a^p\gamma'\int_0^1 \frac{t^p}{(1-t^2)^{(p-r)/2}}\,F\left(\frac{p}{2},\frac{p-r}{2};\frac{p-r}{2}+1;-a^2\frac{t^2}{1-t^2}\right)dt, \tag{4.23}$$

where $\gamma = rc_r c_{n-r-1}/c_n$ and $\gamma' = \Gamma\left((r+1)/2+1\right)/\left((p-r)\Gamma\left(1/2\right)\Gamma\left(r/2\right)\right)$. Consider the integral in (4.23). By setting $z = t^2/(1-t^2)$, we obtain

$$\begin{aligned}&\int_0^1 \frac{t^p}{(1-t^2)^{(p-r)/2}} F\left(\frac{p}{2}, \frac{p-r}{2}; \frac{p-r}{2}+1; -a^2\frac{t^2}{1-t^2}\right) dt \\ &= \frac{1}{2a^{p-r-2}} \int_0^\infty \frac{z^{(p-1)/2}}{(a^2+z)^{(r+3)/2}} F\left(\frac{p}{2}, \frac{p-r}{2}; \frac{p-r}{2}+1; -z\right) dz. \end{aligned} \tag{4.24}$$

We notice that

$$\frac{z^{(p-1)/2}}{(a^2+z)^{(r+3)/2}} \leq \frac{z^{(p-r)/2}}{(a^2+z)^2}.$$

Using this inequality and [1, eq. 7.51, 10], (4.24) can be bounded as follows:

$$\begin{aligned}&\frac{1}{2a^{p-r-2}} \int_0^\infty \frac{z^{(p-1)/2}}{(a^2+z)^{(r+3)/2}} F\left(\frac{p}{2}, \frac{p-r}{2}; \frac{p-r}{2}+1; -z\right) dz \\ &\leq \frac{1}{2a^{p-r-2}} \int_0^\infty \frac{z^{(p-r)/2}}{(a^2+z)^2} F\left(\frac{p}{2}, \frac{p-r}{2}; \frac{p-r}{2}+1; -z\right) dz \\ &= \frac{1}{2a^{p-r-2}} \frac{\Gamma\left((p-r)/2+1\right)\Gamma\left(r/2+1\right)}{\Gamma\left(p/2+1\right)} F\left(\frac{r}{2}+1, 1; \frac{p}{2}+1; 1-a^2\right). \end{aligned} \tag{4.25}$$

Substituting (4.25) in (4.23), we get

$$\begin{aligned}[e_k^{\mathrm{ran}}(A,p)]^{1/p} &\geq a^r \frac{\Gamma\left((p-r)/2\right)\Gamma\left((r+1)/2+1\right)}{\Gamma\left(p/2\right)\Gamma\left(1/2\right)} \\ &- a^{r+2} \frac{r\Gamma\left((r+1)/2+1\right)\Gamma\left((p-r)/2\right)}{\Gamma\left(p/2+1\right)\Gamma\left(1/2\right)} F\left(\frac{r}{2}+1, 1; \frac{p}{2}+1; 1-a^2\right). \end{aligned}$$

This concludes the proof. ∎

4.3. Discussion. Theorems 4.1 and 4.2 state that the randomized error $e_k^{\mathrm{ran}}(A,p)$ must depend on the ratio λ_{r+1}/λ_1. In addition, these theorems describe the actual behavior of the rate of convergence for every k, p, and r. We notice that only when $r > p$ do we have the same rate of convergence as in the asymptotic deterministic case with $\sum_{i=1}^r b_i^2 \neq 0$. For the other two cases, $r = p$ and $r < p$, the rate convergence is slower. This is due to the fact that Theorems 4.1 and 4.2 deal with the randomized case. So, in order to compute the randomized error we have to integrate over all possible starting vectors, even those for which the power method does not converge or converges very slowly.

To give an intuitive idea about the difference in the rate of convergence between the asymptotic deterministic case (the rate is then proportional to $(\lambda_{r+1}/\lambda_1)^k$) and the randomized case, let us analyze the error for $p = 1$. In this case we have only two possibilities: $r > p$ or $r = p = 1$. Assuming $\sum_{i=1}^r b_i^2 \neq 0$, we have

$$\sin(\alpha_k(\mathbf{b})) = \left(\frac{\lambda_{r+1}}{\lambda_1}\right)^k \sqrt{\frac{b_{r+1}^2 + \cdots + b_{r+s}^2}{b_1^2 + \cdots + b_r^2}} + o\left(\left(\frac{\lambda_{r+1}}{\lambda_1}\right)\right),$$

where s is the multiplicity of the second largest eigenvalue.

If $r = 1$, the expected value of $\sin(\alpha_k(\mathbf{b}))$ with respect to $\mathbf{b}$ cannot be proportional to $(\lambda_2/\lambda_1)^k$ since

$$\int_{\|\mathbf{b}\|=1} \sqrt{\frac{b_2^2 + \cdots + b_{s+1}^2}{b_1^2}}\mu(db) = +\infty.$$

A more careful analysis shows that we have to lose a factor proportional to $\ln(\lambda_1/\lambda_2)^{2k}$ in order to achieve the convergence of the integral. For $r \geq 2$,

$$\int_{\|\mathbf{b}\|=1} \sqrt{\frac{b_{r+1}^2 + \cdots + b_{r+s}^2}{b_1^2 + \cdots + b_r^2}}\mu(db) < +\infty,$$

so we have a rate of convergence proportional to $(\lambda_{r+1}/\lambda_1)^k$, as in the deterministic case. The explanation of the general case $p \geq 1$ is similar.

Analyzing upper and lower bounds together, we see the complete behavior of the power method for computing a largest eigenvector. In fact, for every p and r, upper and lower bounds exhibit the same dependence on λ_{r+1}/λ_1 and on k.

We now comment on the bounds proposed by Kostlan in [2].

Kostlan estimates the number of steps required by the power method to give a dominant ε-eigenvector, averaged over all the possible starting vectors. However, he considers another error criterion, so it is not easy to compare these bounds with our bounds. In particular, we use the Euclidean distance, where in [2] the Riemannian distance is considered. Moreover, we study the error in the $\mathcal{L}_p$ case, while Kostlan simply integrates the error over the all possible starting vectors.

5. Asymptotic behavior. In section 4 we provide upper and lower bounds for the randomized error of the power method for each step k. These bounds differ only by multiplicative constants and by lower order terms. We notice that only for upper bounds do the constants depend on the size of the matrix, while for the lower bounds they depend only on p and r. Moreover, if A is a large matrix, the constants of the upper bound become huge. So, it is natural to ask if these constants are sharp. We answer this question by analyzing the asymptotic behavior of the randomized error $e_k^{\mathrm{ran}}(A, p)$.

THEOREM 5.1. *Let A be a symmetric positive definite matrix and let r, $r < n$, and s denote the multiplicities of the two largest eigenvalues λ_1 and λ_{r+1} of A. Then for every p, $1 \leq p < \infty$, we have*

$$\lim_{k\to+\infty} \frac{e_k^{\mathrm{ran}}(A,p)}{x_{r+1}^k} = \left(\frac{\Gamma\left((r-p)/2\right)\Gamma\left((p+s)/2\right)}{\Gamma\left(r/2\right)\Gamma\left(s/2\right)}\right)^{1/p} \qquad \textit{for } p < r,$$

$$\lim_{k\to+\infty} \frac{e_k^{\mathrm{ran}}(A,p)}{x_{r+1}^k\,(2k)^{1/r}\left[\ln\left(1/x_{r+1}\right)\right]^{1/r}} = \left(\frac{\Gamma\left((p+s)/2\right)}{\Gamma\left(p/2\right)\Gamma\left(s/2\right)}\right)^{1/p} \qquad \textit{for } p = r,$$

$$\lim_{k\to+\infty} \frac{e_k^{\mathrm{ran}}(A,p)}{x_{r+1}^{kr/p}} = \left(\frac{\Gamma\left((p-r)/2\right)\Gamma\left((r+s)/2\right)}{\Gamma\left(p/2\right)\Gamma\left(s/2\right)}\right)^{1/p} \qquad \textit{for } p > r.$$

Proof. From (2.9) we have

$$e_k^{\mathrm{ran}}(A,p) = \left(\frac{1}{c_n}\int_{B_n}\left(\frac{\sum_{i=r+1}^n b_i^2 x_i^{2k}}{\sum_{i=1}^r b_i^2 + \sum_{i=r+1}^n b_i^2 x_i^{2k}}\right)^{p/2} db\right)^{1/p}.$$

We can rewrite the previous equation as follows:

$$e_k^{\mathrm{ran}}(A,p) = \left(\frac{1}{c_n} \int_{B_n} \left(\left(\frac{x_{r+1}^{2k} \sum_{i=r+1}^{r+s} b_i^2}{\sum_{i=1}^{r} b_i^2 + x_{r+1}^{2k} \sum_{i=r+1}^{r+s} b_i^2} \right)^{1/2} + r_k(b) \right)^p db \right)^{1/p},$$

where

$$r_k(b) = \left(\frac{\sum_{i=r+1}^{n} b_i^2 x_i^{2k}}{\sum_{i=1}^{r} b_i^2 + \sum_{i=r+1}^{n} b_i^2 x_i^{2k}} \right)^{1/2} - \left(\frac{x_{r+1}^{2k} \sum_{i=r+1}^{r+s} b_i^2}{\sum_{i=1}^{r} b_i^2 + x_{r+1}^{2k} \sum_{i=r+1}^{r+s} b_i^2} \right)^{1/2}. \tag{5.1}$$

Let

$$\tilde{e}_k^{\mathrm{ran}}(A,p) = \left(\frac{1}{c_n} \int_{B_n} \left(\frac{x_{r+1}^{2k} \sum_{i=r+1}^{r+s} b_i^2}{\sum_{i=1}^{r} b_i^2 + x_{r+1}^{2k} \sum_{i=r+1}^{r+s} b_i^2} \right)^{p/2} db \right)^{1/p}.$$

We want to show that

$$\lim_{k \to +\infty} e_k^{\mathrm{ran}}(A,p) = \lim_{k \to +\infty} \tilde{e}_k^{\mathrm{ran}}(A,p). \tag{5.2}$$

Notice that

$$\tilde{e}_k^{\mathrm{ran}}(A,p) \le e_k^{\mathrm{ran}}(A,p) \le \tilde{e}_k^{\mathrm{ran}}(A,p) + \|r_k\|_p,$$

where

$$\|r_k\|_p = \left(\frac{1}{c_n} \int_{B_n} r_k(b)^p \, db \right)^{1/p}.$$

Since $r_k(b) \to 0$ pointwise almost everywhere, and $|r_k(b)| \le 1$ for the $\mathcal{L}_p$-dominated convergence theorem (see [6, p. 312]) we have $\lim_{k \to +\infty} \|r_k\|_p = 0$. This proves (5.2).

Equation (5.2) shows that the asymptotic behavior of $e_k^{\mathrm{ran}}(A,p)$ can be studied by analyzing $\tilde{e}_k^{\mathrm{ran}}(A,p)$. Let $a = x_{r+1}^k$. Integrating with respect to $b_{r+s+1}, \ldots, b_n$, we have

$$[\tilde{e}_k^{\mathrm{ran}}(A,p)]^p = a^p \, \frac{c_{n-r-s}}{c_n} \int_{B_{r+s}} \left(\frac{\sum_{i=r+1}^{r+s} b_i^2}{\sum_{i=1}^{r} b_i^2 + a^2 \sum_{i=r+1}^{r+s} b_i^2} \right)^{p/2} \left(1 - \sum_{i=1}^{r+s} b_i^2 \right)^{(n-r-s)/2} db.$$

Let $\|b\|^2 = \sum_{i=1}^{r} b_i^2$ and let $t_i = b_i/(1 - \|b\|^2)^{1/2}$ for $i = r+1, \ldots, r+s$, and $\|t\|^2 = \sum_{i=r+1}^{r+s} t_i^2$. If we rewrite the last integral as an integral over the balls B_r and B_s, we have

$$[\tilde{e}_k^{\mathrm{ran}}(A,p)]^p = a^p \, \frac{c_{n-r-s}}{c_n} \int_{B_r} \int_{B_s} \frac{\|t\|^p (1 - \|b\|^2)^{(n+p-r)/2} (1 - \|t\|^2)^{(n-r-s)/2}}{[\|b\|^2 + a^2 \|t\|^2 (1 - \|b\|^2)]^{p/2}} dt \, db.$$

Using [1, eq. 4.642] for both integrals, we get

$$\begin{aligned} &[\tilde{e}_k^{\mathrm{ran}}(A,p)]^p \\ &= a^p \gamma \int_0^1 \int_0^1 \frac{t^{s-1} \, b^{r-1} \, t^p (1 - b^2)^{(n+p-r)/2} \, (1 - t^2)^{(n-r-s)/2}}{[b^2 + a^2 t^2 (1 - b^2)]^{p/2}} \, dt \, db \\ &= a^p \gamma \int_0^1 t^{p+s-1} (1 - t^2)^{(n-r-s)/2} \left[\int_0^1 \frac{b^{r-1} (1 - b^2)^{(n+p-r)/2}}{[b^2 + a^2 t^2 (1 - b^2)]^{p/2}} \, db \right] dt, \end{aligned} \tag{5.3}$$

where $\gamma = rsc_{n-r-s}c_rc_s/c_n$.

We have now three cases depending on the relation between p and r. Consider first the case $p < r$. Then the last integral of (5.3) is finite even for $a = 0$. Substituting $a = 0$, we get

$$[\tilde{e}_k^{\mathrm{ran}}(A,p)]^p = a^p\gamma \int_0^1 t^{p+s-1}(1-t^2)^{(n-r-s)/2}\,dt \int_0^1 b^{r-p-1}(1-b^2)^{(n+p-r)/2}\,db.$$

From the definition of the beta function (2.11) we have

$$[\tilde{e}_k^{\mathrm{ran}}(A,p)]^p = a^p\,\frac{\gamma}{4}\,B\left(\frac{p+s}{2}, \frac{n-r-s}{2}+1\right) B\left(\frac{r-p}{2}, \frac{n+p-r}{2}+1\right).$$

Using (2.10), we can express c_i in terms of the gamma function. We obtain

$$[\tilde{e}_k^{\mathrm{ran}}(A,p)]^p = a^p\,\frac{\Gamma\left((r-p)/2\right)\Gamma\left((p+s)/2\right)}{\Gamma\left(r/2\right)\Gamma\left(s/2\right)}.$$

This proves that for $p < r$, by using (5.2) we have

$$\lim_{k\to+\infty} \frac{e_k^{\mathrm{ran}}(A,p)}{x_{r+1}^k} = \left(\frac{\Gamma\left((r-p)/2\right)\Gamma\left((p+s)/2\right)}{\Gamma\left(r/2\right)\Gamma\left(s/2\right)}\right)^{1/p}.$$

Consider now the case $p = r$. From (5.3) we have that $[\tilde{e}_k^{\mathrm{ran}}(A,p)]^p$ is equal to

$$a^p\gamma \int_0^1 t^{p+s-1}(1-t^2)^{(n-p-s)/2}\left[\int_0^1 \frac{b^{p-1}(1-b^2)^{n/2}}{[b^2+a^2t^2(1-b^2)]^{p/2}}\,db\right] dt. \tag{5.4}$$

We expand $b^{p-1}(1-b^2)^{n/2}$ as $b^{p-1}-(n/2)b^{p+1}+O\left(b^{p+3}\right)$. Since $[b^2(1-a^2t^2)+a^2t^2]^{p/2}$ behaves as $b^p+o\left(a^2t^2\right)$, it is sufficient to consider the first two terms of the expansion. As a approaches zero, we have

$$\begin{aligned}
&\int_0^1 \frac{b^{p-1}(1-b^2)^{n/2}}{[b^2(1-a^2t^2)+a^2t^2]^{p/2}}\,db \\
&= \int_0^1 \frac{b^{p-1}}{[b^2(1-a^2t^2)+a^2t^2]^{p/2}}\,db + O\left(\int_0^1 \frac{b^{p+1}}{[b^2(1-a^2t^2)+a^2t^2]^{p/2}}\,db\right) \\
&= \int_0^1 \frac{b^{p-1}}{(b^2+a^2t^2)^{p/2}}\,db + O\left(\int_0^1 b\,db\right).
\end{aligned}$$

Observe that $(b^2+a^2t^2)^{p/2} = b^p + (p/2)b^{2(p/2-1)}a^2t^2(1+o\,(1))$ as $a \to 0$. Then from the last equation we have

$$\begin{aligned}
&\int_0^1 \frac{b^{p-1}}{(b^2+a^2t^2)^{p/2}}\,db + O\left(\int_0^1 b\,db\right) \\
&= \int_0^1 \frac{b^{p-1}}{b^{p-2}\left(b^2+p/2\,a^2t^2\right)}\,db + O\,(1) \\
&= \int_0^1 \frac{b}{b^2+p/2\,a^2t^2}\,db + O\,(1) \\
&= \frac{1}{2}\ln\left(b^2+\frac{p}{2}a^2t^2\right)\Big|_0^1 + O\,(1) \\
&= \ln\left(\sqrt{\frac{2}{pa^2t^2}}\right)(1+o\,(1)).
\end{aligned}$$

Substituting this equality into (5.4) we get

$$[\tilde{e}_k^{\mathrm{ran}}(A,p)]^p = a^p\gamma \int_0^1 t^{p+s-1}(1-t^2)^{(n-p-s)/2} \ln\left(\sqrt{\frac{2}{pa^2t^2}}\right) dt$$
$$= a^p \frac{\gamma}{4} \ln\left(\frac{2}{pa^2}\right) B\left(\frac{p+s}{2}, \frac{n-p-s}{2}+1\right) + O\left(a^p\right).$$

If we replace the expression for γ in the last equation, from (5.2) we obtain

$$\lim_{k\to+\infty} \frac{e_k^{\mathrm{ran}}(A,p)}{x_{r+1}^k (2k)^{1/r} \left[\ln\left(1/x_{r+1}\right)\right]^{1/r}} = \left(\frac{\Gamma\left((p+s)/2\right)}{\Gamma\left(p/2\right)\Gamma\left(s/2\right)}\right)^{1/p}.$$

The last case is $p > r$. We want to compute the limit

$$\lim_{x\to+\infty} \frac{e_k^{\mathrm{ran}}(A,p)}{x_{r+1}^{kr/p}} = \left[\lim_{k\to+\infty} \frac{[\tilde{e}_k^{\mathrm{ran}}(A,p)]^p}{x_{r+1}^{kr}}\right]^{1/p}.$$

From (5.3) we get

$$\lim_{k\to+\infty} \frac{[e_k^{\mathrm{ran}}(A,p)]^p}{x_{r+1}^{kr}} = \lim_{a\to 0} \frac{[e_k^{\mathrm{ran}}(A,p)]^p}{a^r} \tag{5.5}$$
$$= \lim_{a\to 0} a^{p-r}\gamma \int_0^1 t^{p+s-1}(1-t^2)^{(n-r-s)/2} \left[\int_0^1 \frac{b^{r-1}(1-b^2)^{(n+p-r)/2}}{[b^2+a^2t^2(1-b^2)]^{p/2}}\, db\right] dt.$$

Observe that for $a \to 0$ we have

$$\int_0^1 a^{p-r} \frac{b^{r-1}(1-b^2)^{(n+p-r)/2}}{[b^2+a^2t^2(1-b^2)]^{p/2}}\, db$$
$$= \int_0^1 a^{p-r} \frac{b^{r-1}(1-b^2)^{(n+p-r)/2}}{[b^2+a^2t^2]^{p/2}}\, db$$
$$= \int_0^1 a^{p-r} \frac{b^{r-1}(1-b^2)^{(n+p-r)/2}}{a^p t^p \left(b^2/(a^2t^2)+1\right)^{p/2}}\, db. \tag{5.6}$$

We change variables by setting $y = b/(at)$. Then the integral (5.6) becomes

$$\frac{1}{t^{p-r}} \int_0^{1/(at)} \frac{y^{r-1}(1-a^2t^2y^2)^{(n+p-r)/2}}{(y^2+1)^{p/2}}\, dy.$$

If we set $z = y^2$, this integral can be transformed into

$$\frac{1}{2t^{p-r}} \int_0^{1/(a^2t^2)} \frac{z^{r/2-1}(1-a^2t^2z)^{(n+p-r)/2}}{(z+1)^{p/2}}\, dz.$$

We substitute this integral into (5.5). We get

$$\lim_{k\to+\infty} \frac{[\tilde{e}_k^{\mathrm{ran}}(A,p)]^p}{x_{r+1}^{kr}} = \lim_{a\to 0} \frac{[e_k^{\mathrm{ran}}(A,p)]^p}{a^r} \tag{5.7}$$

$$= \frac{\gamma}{2} \int_0^1 t^{r+s-1}(1-t^2)^{(n-r-s)/2} \left[\lim_{a\to 0} \int_0^{1/(a^2t^2)} \frac{z^{r/2-1}(1-a^2t^2z)^{(n+p-r)/2}}{(z+1)^{p/2}}\, dz\right] dt.$$

To find the limit of the last integral, we use the following bounds (for $a < 1$):

$$\int_0^{1/(at)} \frac{z^{r/2-1}(1-at)^{(n+p-r)/2}}{(z+1)^{p/2}} dz \le \int_0^{1/(a^2t^2)} \frac{z^{r/2-1}(1-a^2t^2z)^{(n+p-r)/2}}{(z+1)^{p/2}} dz$$

and

$$\int_0^{1/(a^2t^2)} \frac{z^{r/2-1}(1-a^2t^2z)^{(n+p-r)/2}}{(z+1)^{p/2}} dz \le \int_0^{1/(a^2t^2)} \frac{z^{r/2-1}}{(z+1)^{p/2}}\, dz.$$

Since

$$\lim_{a\to 0} \int_0^{1/(at)} \frac{z^{r/2-1}(1-at)^{(n+p-r)/2}}{(z+1)^{p/2}}\, dz = \lim_{a\to 0} \int_0^{1/(a^2t^2)} \frac{z^{r/2-1}}{(z+1)^{p/2}}\, dz,$$

passing to the limit and then using [1, eq. 3.194, 3], we get

$$\lim_{a\to 0} \int_0^{1/(a^2t^2)} \frac{z^{r/2-1}}{(z+1)^{p/2}}\, dz = \int_0^{+\infty} \frac{z^{r/2-1}}{(z+1)^{p/2}}\, dz = B\left(\frac{r}{2}, \frac{p-r}{2}\right).$$

Hence, we also have

$$\lim_{a\to 0} \int_0^{1/(a^2t^2)} \frac{z^{r/2-1}(1-a^2t^2z)^{(n+p-r)/2}}{(z+1)^{p/2}}\, dz = B\left(\frac{r}{2}, \frac{p-r}{2}\right).$$

From (5.7), we get

$$\begin{aligned}
&\lim_{k\to+\infty} \frac{[e_k^{\mathrm{ran}}(A,p)]^p}{x_{r+1}^{kr}} \\
&= \frac{\gamma}{2}\, B\left(\frac{r}{2}, \frac{p-r}{2}\right) \int_0^1 t^{r+s-1}(1-t^2)^{(n-r-s)/2} dt \\
&= \frac{\Gamma\left((p-r)/2\right)\, \Gamma\left((r+s)/2\right)}{\Gamma\left(p/2\right) \Gamma\left(s/2\right)}.
\end{aligned}$$

This concludes the proof. ∎

Theorem 5.1 shows that upper and lower bounds provided in section 4 are asymptotically optimal. In fact, the analysis of the asymptotic case indicates that the upper and lower bounds cannot be improved since the constants coincide with those of the upper bound when we set the multiplicity of the second largest eigenvalue to $n - r$, and with those of the lower bound for $s = 1$. The constants increase with s and $1/r$. This corresponds to the intuitive idea that the convergence is fast if the eigenspace $\mathcal{Z}$ is large and is slow if the eigenspace corresponding to the second largest eigenvalue is large. Note that if p approaches infinity, the rate of convergence approaches 1 and even the constant converges to 1. This agrees with (2.8) for $p = \infty$.

6. Numerical tests. We tested the power method for several matrices with many pseudorandom starting vectors $\mathbf{b}$. The matrix A can be chosen as follows. As before, let $\mathbf{u}_k(A, \mathbf{b})$ be the vector computed by the power method applied to the matrix A with starting vector $\mathbf{b}$. Observe that for any orthogonal matrix Q, we have $\mathbf{u}_k(Q^TAQ, Q^T\mathbf{b}) = \mathbf{u}_k(A, \mathbf{b})$. Moreover, the uniform distribution on the unit sphere of the vectors $\mathbf{b}$ implies the same distribution of vectors $Q^T\mathbf{b}$. So, without

TABLE 6.1
Quadratic distribution 2 with the eigenvalues $\lambda_i = 2(1-(i/101)^2)$.

k	$\varepsilon^{\mathrm{ran}}$	$\varepsilon^{\mathrm{worst}}$	$\varepsilon^{\mathrm{best}}$	$\varepsilon^{\mathrm{lb}}$	$\varepsilon^{\mathrm{ub}}$	p
10	$9.737e-01$	$9.999e-01$	$7.567e-01$	$4.782e-01$	$7.998e+00$	1
100	$9.111e-01$	$9.999e-01$	$4.149e-01$	$4.850e-01$	$7.992e+00$	1
1000	$7.114e-01$	$9.999e-01$	$6.811e-02$	$5.185e-01$	$7.685e+00$	1
10	$9.735e-01$	$9.999e-01$	$7.226e-01$	$6.457e-01$	$3.522e+00$	2
100	$9.239e-01$	$9.999e-01$	$3.319e-01$	$6.394e-01$	$3.474e+00$	2
1000	$7.383e-01$	$9.999e-01$	$7.003e-02$	$5.799e-01$	$3.035e+00$	2
10	$9.779e-01$	$1.000e+00$	$7.649e-01$	$2.712e-01$	$1.129e+00$	10
100	$9.412e-01$	$9.999e-01$	$3.882e-01$	$2.729e-01$	$1.127e+00$	10
1000	$8.675e-01$	$1.000e-01$	$5.303e-02$	$2.902e-01$	$1.097e+00$	10

loss of generality, we can restrict ourselves only to considering diagonal matrices; see also [4] and [5]. Vectors uniformly distributed over the unit sphere can be generated as described in [4] and [5].

The tests were performed on a Sun SPARCsystem 10 using double precision. To compute the values of the hypergeometric and gamma functions we used the program *Mathematica*.

We tested many different matrices of size 100 with the distributions of the eigenvalues chosen as in [5]. We tested the following distributions:

- Chebyshev distribution: $\lambda_i = 1 + \cos(((2i-1)\pi)/200)$;
- quadratic distribution 1: $\lambda_i = 2\,(1-i/101)^2$;
- quadratic distribution 2: $\lambda_i = 2(1-(i/101)^2)$;
- uniform distribution: $\lambda_i = 2(1-i/101)$;
- logarithmic distribution: $\lambda_i = 2\log(102-i)/\log(102)$;
- exponential distribution 1: $\lambda_i = 2\,e^{-\sqrt[3]{i}}$;
- exponential distribution 2: $\lambda_i = 1 + e^{-i}$.

From the theoretical bounds (see Theorems 4.1 and 4.2), it turns out that the behavior of the power method depends on the relation between r and p. We tested the power method for different values of p and r for a fixed ratio between the two largest eigenvalues.

The main goal of these tests was to verify the results proved in Theorems 4.1 and 4.2 and to see how much upper and lower bounds differ from the experimental values.

In order to approximate the randomized error $e_k^{\mathrm{ran}}(A,p)$ we have used 1,000 pseudorandom vectors $\mathbf{b}$. So, the randomized error is replaced by $\varepsilon^{\mathrm{ran}}$ obtained as the mean value among the 1,000 pseudorandom vectors, i.e.,

$$\varepsilon^{\mathrm{ran}} = \left(\frac{1}{1,000} \sum_{i=1}^{1,000} \sin^p(\alpha_k(b_i)) \right)^{1/p}.$$

By $\varepsilon^{\mathrm{worst}}$ and $\varepsilon^{\mathrm{best}}$ we denote, respectively, the worst and best value of $\sin(\alpha_k(b_i))$. These values give an indication about how much $\varepsilon^{\mathrm{ran}}$ differs from the values $\sin(\alpha_k(b_i))$. Let $\varepsilon^{\mathrm{lb}}$ and $\varepsilon^{\mathrm{ub}}$ denote the lower and the upper bounds computed using formulas given by Theorems 4.2 and 4.1. Finally, k and p are the number of iterations and the parameter of the norm, respectively.

In order to underline the dependence of the rate of convergence on the ratio between the two largest eigenvalues we report the results obtained for the quadratic distribution 2 (see Table 6.1) and the exponential distribution 1 (see Table 6.2). In

TABLE 6.2
Exponential distribution 1 *with the eigenvalues* $\lambda_i = 2\,e^{-i^{1/3}}$.

k	ε^{ran}	$\varepsilon^{\text{worst}}$	$\varepsilon^{\text{best}}$	ε^{lb}	ε^{ub}	p
10	$1.770e-01$	$9.999e-01$	$9.630e-04$	$1.698e-01$	$2.124e+00$	1
30	$2.432e-03$	$8.996e-01$	$1.077e-06$	$2.300e-03$	$2.864e-02$	1
10	$2.509e-01$	$9.999e-01$	$7.056e-04$	$2.368e-01$	$9.616e-01$	2
30	$2.468e-02$	$7.652e-01$	$9.823e-07$	$2.006e-02$	$7.148e-02$	2
10	$6.801e-01$	$9.999e-01$	$2.079e-03$	$3.888e-01$	$8.715e-01$	10
30	$3.562e-01$	$7.081e-01$	$1.977e-06$	$3.421e-01$	$5.182e-01$	10

TABLE 6.3
Modified exponential distribution 2 *with the eigenvalues* $\lambda_1 = \lambda_2 = 1+e^{-1}$ *and* $\lambda_i = 1+e^{-(i-1)}$ *for* $i = 3, \ldots, n$.

k	ε^{ran}	$\varepsilon^{\text{worst}}$	$\varepsilon^{\text{best}}$	ε^{lb}	ε^{ub}	p
10	$4.100e-01$	$9.960e-01$	$1.124e-01$	$1.276e-01$	$1.920e+00$	1
30	$3.765e-03$	$1.075e-01$	$2.915e-04$	$3.693e-03$	$4.622e-02$	1
10	$4.593e-01$	$9.989e-01$	$1.171e-01$	$1.570e-01$	$3.650e+00$	2
30	$7.979e-03$	$1.063e-01$	$2.693e-04$	$7.511e-03$	$1.245e-01$	2
10	$6.904e-01$	$9.973e-01$	$1.190e-01$	$2.472e-01$	$8.850e-01$	10
30	$2.551e-01$	$5.089e-01$	$2.435e-04$	$1.950e-01$	$4.200e-01$	10

fact, these distributions are those (among the different distributions considered) for which we have the largest (the smallest) ratio between λ_2 and λ_1 and then the slowest (the fastest) convergence, respectively.

From Table 6.1 we see that for three different values of p, even after 1,000 iterations the randomized error is still very close to 1. An important observation concerns the lower and upper bounds. We notice that the lower bound is a good approximation of the expected value ε^{ran} while the upper bound is clearly an overestimate. This is due to the following reasons.

1. The constants in the upper bounds (see Theorem 4.1) grow with the size of the matrix.
2. Since the ratio $x_2 = \lambda_2/\lambda_1$ is very close to 1, x_2^k goes very slowly to 0 with k. In this case, the upper bound is more sensitive to the big multiplicative constants.

Table 6.2 is more interesting since it allows us to see the dependence of the speed of convergence on p and r. The speed of convergence is now good. In fact, after only 30 iterations we get an error of the order of 10^{-3} when $p = r = 1$. In this case, we have also that ε^{lb} and ε^{ub} are relatively close to each other and that the error ε^{ran} for $k = 30$ is very close to the theoretical lower bound.

In general, it is possible to observe that the values of ε^{ran} computed with these tests are very close to the theoretical lower bounds, while they are more distant from the upper bounds even for small λ_{r+1}/λ_1. This is due to the importance of the multiplicity s of λ_{r+1}, which results from the asymptotic constants of Theorem 5.1. Experimental results prove that the power method behaves differently for matrices with the same two largest eigenvalues but with different multiplicities. In particular, increasing s, we get bounds closer to the upper bounds.

To understand the role of p and r, we have performed tests with matrices for which the multiplicity of the largest eigenvalue is $r \geq 2$. In Table 6.3 we report the results for the modified exponential distribution 2 with $r = 2$.

An important observation concerns the comparison between the three cases, $p < r$,

$p = r$, and $p > r$. From Table 6.3 it is easy to see that for the same value of k, the rates of convergence are different. For example, for $k = 30$ we have an error of the order of 10^{-3} for $p \leq r$, and of order 10^{-1} for $p > r$.

We also performed tests with matrices with only two distinct eigenvalues. These tests indicate the asymptotic dependence of the randomized error on the multiplicity s of the second eigenvalue. In particular, they show that ε^{ran} is closer to ε^{ub} when s is big. This is an important consequence of Theorem 5.1.

7. Conclusions. In this paper we have investigated the convergence of the power method for approximating an eigenvector corresponding to the largest eigenvalue. As our error measure, we have taken the sine of the acute angle $\alpha_k(\mathbf{b})$ between the vector computed by the power method after k steps with the starting vector $\mathbf{b}$, and the eigenspace related to the largest eigenvalue. We have analyzed the $\mathcal{L}_p$ norm of $\sin(\alpha_k(\cdot))$ for $p \in [1, +\infty]$. We have shown that, if the starting vector $\mathbf{b}$ is chosen according to the uniform distribution over the unit sphere, the rate of convergence depends on the ratio between the two largest eigenvalues. In particular, if r is the multiplicity of the largest eigenvalue λ_1, and the $\mathcal{L}_p$ norm is used, then the randomized error is proportional to $(\lambda_{r+1}/\lambda_1)^k$ if $p < r$, to $(\lambda_{r+1}/\lambda_1)^{kr/p}$ if $p > r$, and to $k^{1/p}\,(\lambda_{r+1}/\lambda_1)^k$ if $p = r$.

For every $p \in [1, +\infty)$, we have found asymptotic and nonasymptotic bounds, and we have shown that the asymptotic constants are equal to those obtained for the upper and lower bounds when the multiplicity of the second largest eigenvalue is set to $n - r$ and 1, respectively. We stress that our results hold for a class of norms and that they show how, by using a different norm, we can have a different speed of convergence. Our bounds depend on the distribution of the eigenvalues, and we have proven that this is unavoidable. Comparing with results of [4], we conclude that approximating a largest eigenvector by the power method is more difficult than approximating the largest eigenvalue in the randomized setting.

Acknowledgment. I wish to thank Henryk Woźniakowski for the guidance and valuable help he provided during all the stages of this work.

REFERENCES

[1] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, New York, 1994.
[2] E. Kostlan, *Complexity theory for numerical linear algebra*, J. Comput. Appl. Math., 22 (1988), pp. 219–230.
[3] E. Kostlan, *Statistical complexity of dominant eigenvector calculation*, J. Complexity, 7 (1991), pp. 371–379.
[4] J. Kuczyński and H. Woźniakowski, *Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1094–1122.
[5] J. Kuczyński and H. Woźniakowski, *Probabilistic bounds on the extremal eigenvalues and condition number by the Lanczos algorithm*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 672–691.
[6] S. Lang, *Analysis* II, Addison–Wesley, Reading, MA, 1969.
[7] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[8] M. Shub, *The geometry and topology of dynamical systems and algorithms for numerical problems*, in Proc. of the 1983 Beijing Symposium on Differential Geometry and Differential Equations, Liao Shantao, ed., Science Press, Beijing, China, 1986.
[9] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.
[10] P. E. Wright, *Statistical complexity of the power method for Markov chains*, J. Complexity, 5 (1989), pp. 119–143.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 4, pp. 938–958, October 1997

010

ON THE STABILITY OF NULL-SPACE METHODS FOR KKT SYSTEMS*

ROGER FLETCHER† AND TOM JOHNSON†

Abstract. This paper considers the numerical stability of null-space methods for Karush–Kuhn–Tucker (KKT) systems, particularly in the context of quadratic programming. The methods we consider are based on the direct elimination of variables, which is attractive for solving large sparse systems. Ill-conditioning in a certain submatrix A in the system is shown to adversely affect the method insofar as it is commonly implemented. In particular, it can cause growth in the residual error of the solution, which would not normally occur if Gaussian elimination or related methods were used. The mechanism of this error growth is studied and is not due to growth in the null-space basis matrix Z, as might have been expected, but to the indeterminacy of this matrix. When LU factors of A are available it is shown that an alternative form of the method is available which avoids this residual error growth. These conclusions are supported by error analysis and Matlab experiments on some extremely ill-conditioned test problems. These indicate that the alternative method is very robust in regard to residual error growth and is unlikely to be significantly inferior to the methods based on an orthogonal basis matrix. The paper concludes with some discussion of what needs to be done when LU factors are not available.

Key words. null-space method, KKT system, ill-conditioning

AMS subject classifications. 65F05, 65G05

PII. S0895479896297732

1. Introduction.

A KKT system is a linear system

$$\begin{bmatrix} G & A \\ A^T & 0 \end{bmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{c} \\ \mathbf{b} \end{pmatrix} \tag{1.1}$$

involving a symmetric matrix of the form

$$K = \begin{bmatrix} G & A \\ A^T & 0 \end{bmatrix}. \tag{1.2}$$

Such systems are characteristic of the optimization problem

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\mathbf{x}^T G\mathbf{x} - \mathbf{c}^T\mathbf{x} \\ &\text{subject to} && A^T\mathbf{x} = \mathbf{b}, \end{aligned} \tag{1.3}$$

in which there are linear equality constraints and the objective is a quadratic function. The KKT system (1.1) represents the first-order necessary conditions for a solution of this problem, and $\mathbf{y}$ is a vector of Lagrange multipliers (see [6], for example). Problems like (1.3) arise in many fields of study, such as in Newton's method for nonlinear programming (e.g., [6]) and in the solution of partial differential equations involving incompressible fluid flows, incompressible solids, and the analysis of plates and shells (e.g., Bathe [1], Brezzi and Fortin [2]). Also, problems with inequality

*Received by the editors January 22, 1996; accepted for publication (in revised form) by N. J. Higham October 18, 1996. An early version of this paper was presented at the Dundee Biennial Conference in Numerical Analysis, June 1995 and the Manchester IMA Conference on Linear Algebra and Its Applications, July 1995.

http://www.siam.org/journals/simax/18-4/29773.html

†Department of Mathematics and Computer Science, University of Dundee, Dundee DD1 4HN, Scotland, UK (fletcher@mcs.dundee.ac.uk).

constraints are often solved by solving a sequence of equality constrained problems, most particularly in the active set method for quadratic programming.

In (1.2) and (1.3), G is the symmetric $n \times n$ Hessian matrix of the objective function, A is the $n \times m$ Jacobian matrix of the linear constraints, and $m \leq n$. We assume that A has full rank, for otherwise K would be singular. In some applications, A does not immediately have full rank but can readily be reduced to a full rank matrix by a suitable transformation.

There are various ways of solving KKT systems, most of which can be regarded as symmetry-preserving variants of Gaussian elimination with pivoting (see, for example, Gould [9]). This approach is suitable for a one-off solution of a large sparse KKT system, by incorporating a suitable data structure which permits fill-in in the resulting factors. Our interest in KKT systems arises in a quadratic programming (QP) context, where we are using the so-called *null-space method* to solve the sequence of equality constrained problems that arise. This method is described in section 2. An important feature of QP is that the successive matrices K differ only in that one column is either added to or removed from A. The null-space method allows this feature to be used advantageously to update factors of the reduced Hessian matrix that arises when solving the KKT system. However, in this paper we do not consider the updating issue but concentrate on the solution of a single problem like (1.3), but in a null-space context. In fact the null-space method is related to one of the above-mentioned variants of Gaussian elimination, and this point is discussed towards the end of section 3.

In this paper we study the numerical stability of the null-space method when the matrix K is ill conditioned. This arises either when the matrix A is close to being rank deficient or when the reduced Hessian matrix is ill conditioned. It is well known, however, that Gaussian elimination with pivoting usually enables ill conditioned systems to be solved with small backward error (that is, the computed solution is the exact solution of a nearby problem). As Wilkinson [14, Chapter 4] points out, the size of the backward error depends only on the growth in certain reduced matrices, and for an ill-conditioned matrix it is usual for the reduced matrices to diminish in size rather than grow. Although it is possible for exponential growth to occur (we give an example for a KKT system), this is most unlikely in practice. A consequence of this is that if the exact solution is of moderate size, then a very small residual error is obtained from the computed solution. Thus variants of Gaussian elimination with pivoting usually provide a very stable method for solving ill-conditioned systems.

However, this argument does not carry over to the null-space method and we indicate at the end of section 2 that there are serious concerns about numerical stability when A is nearly rank deficient. We describe some Matlab experiments in section 6 which support these concerns. In particular, the residual of the KKT system is seen to be proportional to the condition number of A. We present some error analysis in section 4 which shows how this arises.

When LU factors of A are available, we show in section 3 that there is an alternative way of implementing a null-space method, which avoids the numerical instability. This is also supported by Matlab experiments. The reasons for this are described in section 5, and we present some error analysis which illustrates the difference in the two approaches. In practice, when solving large sparse QP problems, LU factors are not usually available, and it is more usual to use some sort of product form method. We conclude with some remarks about what can be done in this situation to avoid numerical instability.

2. Null-space methods. A null-space method (see, e.g., [6]) is an important technique for solving QP problems with equality constraints. In this section we show how the method can be derived as a generalized form of constraint elimination. The key issue in this procedure is the formation of a basis for the column null space of A. We determine the basis in such a way that we are able to solve large sparse problems efficiently. When A is ill conditioned we argue that there is serious concern for the numerical stability of the method.

The column null space of A may be defined by

$$\mathcal{N}(A) = \{\mathbf{z} \,|\, A^T\mathbf{z} = \mathbf{0}\}$$

and has dimension $n - m$ when A has full rank. Any matrix

$$Z = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{n-m}]$$

whose columns are a basis for $\mathcal{N}(A)$ will be referred to as a *null-space matrix* for A. Such a matrix satisfies $A^TZ = 0$ and has linearly independent columns. A general specification for computing a null-space matrix is to choose an $n \times (n-m)$ matrix V such that the matrix

$$\mathbf{A} = [\, A \quad V \,]$$

is nonsingular. Its inverse is then partitioned in the following way:

$$\mathbf{A}^{-1} = [\, A \quad V \,]^{-1} = \begin{bmatrix} Y^T \\ Z^T \end{bmatrix} \begin{matrix} m \\ n-m \end{matrix}. \tag{2.1}$$

It follows from the properties of the inverse that $A^TZ = 0$ and $A^TY = I_m$. By construction, the columns of Z are linearly independent, and it follows that these columns form a basis for $\mathcal{N}(A)$.

The value of this construction is that it enables us to parametrize the solution set of the (usually) underdetermined system $A^T\mathbf{x} = \mathbf{b}$ in (1.3) by

$$\mathbf{x} = Y\mathbf{b} + Z\mathbf{v}, \qquad \mathbf{v} \in \mathbb{R}^{n-m}. \tag{2.2}$$

Here $Y\mathbf{b}$ is one particular solution of $A^T\mathbf{x} = \mathbf{b}$ and any other solution $\mathbf{x}$ differs from $Y\mathbf{b}$ by a vector, $Z\mathbf{v}$ say, in $\mathcal{N}(\mathcal{A})$. Thus (2.2) provides a general way of eliminating the constraints, by expressing the problem in terms of the *reduced variables* $\mathbf{v}$. Hence, if (2.2) is substituted into the objective function of (1.3), we obtain the *reduced problem*

$$\text{minimize} \qquad \tfrac{1}{2}\mathbf{v}^T(Z^TGZ)\mathbf{v} + \mathbf{v}^TZ^T(GY\mathbf{b} - \mathbf{c}). \tag{2.3}$$

We refer to Z^TGZ as the *reduced Hessian matrix* and $Z^T(GY\mathbf{b} - \mathbf{c})$ as the *reduced gradient vector* (at the point $\mathbf{x} = Y\mathbf{b}$). A necessary and sufficient condition for (2.3) to have a unique minimizer is that Z^TGZ is positive definite. In this case there exist Choleski factors $Z^TGZ = LL^T$, and (2.3) can be solved by finding a stationary point, that is, by solving the linear system

$$LL^T\mathbf{v} = Z^T(\mathbf{c} - GY\mathbf{b}). \tag{2.4}$$

Then substitution of $\mathbf{v}$ into (2.2) determines the solution $\mathbf{x}$ of (1.3). The vector $G\mathbf{x} - \mathbf{c}$ is the gradient of the objective function at the solution, so a vector $\mathbf{y}$ of Lagrange multipliers satisfying $G\mathbf{x} - \mathbf{c} + A\mathbf{y} = \mathbf{0}$ can then be obtained from

$$\mathbf{y} = Y^T(\mathbf{c} - G\mathbf{x}) \tag{2.5}$$

by virtue of the property that $Y^T A = I$. The vectors $\mathbf{x}$ and $\mathbf{y}$ also provide the solution to (1.1), as can readily be verified.

In practice, when A is a large sparse matrix, the matrices Y and Z are usually substantially dense, and it is impracticable to store them explicitly. Instead, products with Y and Z or their transposes are obtained by solving linear systems involving $\mathbf{A}$. For example, the vector $\mathbf{x} = Y\mathbf{b} + Z\mathbf{v}$ in (2.2) could be computed by solving the linear system

$$\mathbf{A}^T\mathbf{x} = \begin{pmatrix} \mathbf{b} \\ \mathbf{v} \end{pmatrix} \tag{2.6}$$

by virtue of (2.1). Likewise, solving the system

$$\mathbf{A}\begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} = \mathbf{t} \tag{2.7}$$

provides the products $\mathbf{u}_1 = Y^T\mathbf{t}$ and $\mathbf{u}_2 = Z^T\mathbf{t}$. These computations require an invertible representation of the matrix $\mathbf{A}$ to be available.

Solving systems involving $\mathbf{A}$ is usually a major cost with the null-space method. To keep this cost as low as possible, it is preferable to choose the matrix V to be sparse. Other choices (for example, based on the QR factors of A; see [6]) usually involve significantly more fill-in and computational expense. In particular, it is attractive to choose the columns of V to be unit vectors, using some form of pivoting to keep $\mathbf{A}$ well conditioned. In this case, assuming that the row permutation has been incorporated into A, it is possible to write

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}, \qquad V = \begin{bmatrix} 0 \\ I \end{bmatrix}, \tag{2.8}$$

where A_1 is an $m \times m$ nonsingular submatrix. Then (2.1) becomes

$$\begin{bmatrix} Y^T \\ Z^T \end{bmatrix} = \begin{bmatrix} A_1 & \\ A_2 & I \end{bmatrix}^{-1} = \begin{bmatrix} A_1^{-1} & \\ -A_2A_1^{-1} & I \end{bmatrix}$$

and provides an explicit expression for Y and Z. In particular, we see that

$$Z^T = \begin{bmatrix} -A_2A_1^{-1} & I \end{bmatrix}. \tag{2.9}$$

We refer to this choice of V as *direct elimination*, as it corresponds to directly using the first m variables to eliminate the constraints (see [6]). We shall adopt this choice of V throughout the rest of the paper.

The reduced Hessian matrix Z^TGZ is also needed for use in (2.3) and can be calculated in a similar way. The method is to compute the vectors $Z^TGZ\mathbf{e}_k$ for $k = 1, 2, \ldots, n-m$, where $\mathbf{e}_k$ denotes column k of the unit matrix I_{n-m}. The computation is carried out from right to left by first computing the vector $\mathbf{z}_k = Z\mathbf{e}_k$ by solving the system

$$\mathbf{A}^T\mathbf{z}_k = \begin{pmatrix} \mathbf{0} \\ \mathbf{e}_k \end{pmatrix}. \tag{2.10}$$

Then the product $G\mathbf{z}_k$ is computed, followed by the solution of

$$\mathbf{A}\mathbf{u} = G\mathbf{z}_k. \tag{2.11}$$

The partition $\mathbf{u}_2$ is then column k of Z^TGZ, as required. The lower triangle of Z^TGZ is then used to calculate the Choleski factor L. A similar approach is essentially used in an active set method for QP, in which the Choleski factor of Z^TGZ is built up over a sequence of iterations. (If indefinite equality constraint QP problems are solved, it may be necessary to solve KKT systems in which Z^TGZ is indefinite. We note that such systems can also be solved in a numerically stable way which preserves symmetry; see Bunch and Parlett [4], Fletcher [5], Bunch and Kaufman [3], and the important recent contribution of Higham [10]. For inequality QP, the indefinite case may be avoided by computing a negative curvature search direction (e.g., Forsgren and Murray [7]). For the purposes of our paper it is immaterial how systems involving Z^TGZ are solved, as long as the method is backward stable.)

An advantage of the null-space approach is that we only need to have available a subroutine or code for the matrix product $G\mathbf{v}$. Thus we can take full advantage of sparsity or structure in G, without, for example, having to allow for fill-in as Gaussian elimination would require. The approach is most convenient when Z^TGZ is sufficiently small to allow it to be stored as a dense matrix. In fact there is a close relationship between the null-space method and a variant of Gaussian elimination, as we shall see in the next section, and the matrix Z^TGZ is the same submatrix in both methods. Thus it would be equally easy (or difficult) to represent Z^TGZ in a sparse matrix format with either method.

To summarize the content of this section we can enumerate the steps implied by (2.2) through (2.5):

1. Calculate Z^TGZ as in (2.10) and (2.11).
2. Calculate $\mathbf{s} = Y\mathbf{b}$ by a solve with $\mathbf{A}^T$ as in (2.6) with $\mathbf{v} = \mathbf{0}$.
3. Calculate $\mathbf{t} = \mathbf{c} - G\mathbf{s}$ requiring a product with G.
4. Calculate $\mathbf{u}_2 = Z^T\mathbf{t}$ by a solve with $\mathbf{A}$ as in (2.7).
5. Solve $Z^TGZ\mathbf{v} = \mathbf{u}_2$ to determine $\mathbf{v}$ as in (2.4).
6. Calculate $\mathbf{x} = Y\mathbf{b} + Z\mathbf{v}$ by a solve with $\mathbf{A}^T$ as in (2.6).
7. Calculate $\mathbf{g} = \mathbf{c} - G\mathbf{x}$ requiring a product with G.
8. Calculate $\mathbf{y} = Y^T\mathbf{g}$ by a solve with $\mathbf{A}$, which also provides $\mathbf{z} = Z^T\mathbf{g}$.

When direct elimination based on (2.9) is used, we shall refer to this as *Method* 1. Step 1 requires $2(n-m)$ solves with either $\mathbf{A}$ or $\mathbf{A}^T$ and $n-m$ products with G to set up the reduced Hessian matrix. The remaining steps require four solves and two products, plus a solve with Z^TGZ. In some circumstances these counts can be reduced. If $\mathbf{b} = \mathbf{0}$ then steps 2 and 3 are not required. If the multiplier part $\mathbf{y}$ of the solution is not of interest, then steps 7 and 8 are not needed. Note that Method 1 does not necessarily require LU factors of $\mathbf{A}$ and may (in a QP code, for example) be implemented using a product form (e.g., Suhl [12]) or a Schur complement (e.g., Gill et al. [8]) representation of $\mathbf{A}$.

We now turn to the concerns about the numerical stability of the null-space method when A (and hence A_1 and $\mathbf{A}$) is ill conditioned. In this case A is close to a rank deficient matrix, A' say, which has a null space of higher dimension. When we solve systems like (2.10) and (2.11), the matrix Z that we are implicitly using is badly determined. Therefore, because of round-off error, we effectively get a significantly different Z matrix each time we carry out a solve. Thus the computed reduced Hessian matrix Z^TGZ does not correspond to any one particular Z matrix. As we shall see in the rest of the paper, this can lead to solutions with significant residual error.

3. Using LU factors of A. In this section we consider the possibility that we can readily compute LU factors of A given by

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} U, \tag{3.1}$$

where L_1 is unit lower triangular and U is upper triangular. We can assume that a row permutation has been made which enables us to bound the elements of L_1 and L_2 by $|l_{ij}| \leq 1$. As we shall see, these factors permit us to circumvent the difficulties caused by ill conditioning to a large extent. (Unfortunately, LU factors are not always available, and some indication is given in section 7 as to what might be done in this situation.) We also describe how the steps in the null-space method are changed. Finally, we explore some connections with Gaussian elimination and other methods, which provide some insight into the likelihood of growth in Z.

A key observation is that if LU factors of A are available, then it is possible to express Z in the alternative form

$$Z^T = [\, -L_2 L_1^{-1} \quad I \,], \tag{3.2}$$

in which the UU^{-1} factors arising from (2.9) and (3.1) are canceled out. A minor disadvantage, compared to (2.9), is that L_2 is needed, which is likely to be less sparse than A_2 and also requires additional storage. However, if A is ill conditioned, this is manifested in U (but not usually L) being ill conditioned, so (3.2) usually enables Z to be defined in a way which is well conditioned. In calculating the reduced Hessian matrix it is convenient to define

$$\mathbf{L} = \begin{bmatrix} L_1 & \\ L_2 & I \end{bmatrix} \tag{3.3}$$

and replace equations (2.10) and (2.11) by

$$\mathbf{L}^T \mathbf{z}_k = \begin{bmatrix} L_1^T & L_2^T \\ & I \end{bmatrix} \mathbf{z}_k = \begin{pmatrix} \mathbf{0} \\ \mathbf{e}_k \end{pmatrix} \tag{3.4}$$

and

$$\mathbf{L}\mathbf{u} = \begin{bmatrix} L_1 & \\ L_2 & I \end{bmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} = G\mathbf{z}_k. \tag{3.5}$$

The steps of the resulting null-space method are as follows (using subscript 1 to denote the first m rows of a vector or matrix and subscript 2 to denote the last $n-m$ rows).

1. Calculate $Z^T G Z$ as in (3.4) and (3.5).
2. Calculate $\mathbf{s}_1 = L_1^{-T} U^{-T} \mathbf{b}$ and let $\mathbf{s} = \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{0} \end{pmatrix}$.
3. Calculate $\mathbf{t} = \mathbf{c} - G\mathbf{s}$ requiring a product with G.
4. Calculate $\mathbf{u}_2 = Z^T \mathbf{t} = \mathbf{t}_2 - L_2 L_1^{-1} \mathbf{t}_1$.
5. Solve $Z^T G Z \mathbf{v} = \mathbf{u}_2$ for $\mathbf{v}$.
6. Calculate $\mathbf{w} = Z\mathbf{v} = \begin{pmatrix} -L_1^{-T} L_2^T \mathbf{v} \\ \mathbf{v} \end{pmatrix}$.
7. Calculate $\mathbf{x} = \mathbf{s} + \mathbf{w}$.
8. Calculate $\mathbf{g} = \mathbf{c} - G\mathbf{x}$ requiring a product with G.
9. Calculate $\mathbf{y} = U^{-1} L_1^{-1} \mathbf{g}_1$.
10. Calculate $\mathbf{z} = Z^T \mathbf{g} = \mathbf{g}_2 - L_2 L_1^{-1} \mathbf{g}_1$.

In the above, inverse operations involving L_1 and U are done by forward or backward substitution. The method is referred to as *Method* 2 in what follows. (For comparability with Method 1, we have also included the calculation of the reduced gradient $\mathbf{z}$, although this would not normally be required.) Note that all solves with the $n \times n$ matrix $\mathbf{A}$ are replaced by solves with smaller $m \times m$ matrices. Also, steps 1, 4, 6, and 10 use the alternative definition (3.2) of Z and so avoid a potentially ill-conditioned calculation with $\mathbf{A}$ (or A_1). We consider the numerical stability of both Method 1 and Method 2 in more detail in the next section.

In the rest of this section, we explore some connections between this method and some variants of Gaussian elimination, and we examine the factored forms that are provided by these methods. It is readily observed (but not well known) that there are block factors of K corresponding to any null-space method in this general format. These are the factors

$$K = \begin{bmatrix} A & V & \\ & & I \end{bmatrix} \begin{bmatrix} Y^TGY & Y^TGZ & I \\ Z^TGY & Z^TGZ & \\ I & & \end{bmatrix} \begin{bmatrix} A^T & & \\ V^T & & \\ & I & \end{bmatrix} \tag{3.6}$$

(using blanks to denote a zero matrix). This result is readily verified by using the equation $AY^T + VZ^T = I$ derived from (2.1). Equation (3.6) makes it clear that if these factors of K are used to solve (1.1), then inverse representations of the matrices $\mathbf{A}$ and Z^TGZ will be required. However, these factors are not directly useful as a method of solution, as they also involve the matrices Y^TGY and Y^TGZ whose computation we wish to avoid in a null-space method. Equation (3.6) also shows that K^{-1} will become large when either $\mathbf{A}$ or Z^TGZ is ill conditioned, and we would expect the spectral condition number of K to behave like $\kappa_K \sim \kappa_{\mathbf{A}}^2 \kappa_M$ where $M = Z^TGZ$.

When using direct elimination (2.8) we may partition K in the form

$$K = \begin{pmatrix} G_{11} & G_{12} & A_1 \\ G_{21} & G_{22} & A_2 \\ A_1^T & A_2^T & 0 \end{pmatrix}.$$

When A has LU factors (3.1) then it is readily verified that another way of factorizing K is given by

(3.7)

$$\begin{bmatrix} G_{11} & G_{12} & A_1 \\ G_{21} & G_{22} & A_2 \\ A_1^T & A_2^T & 0 \end{bmatrix} = \begin{bmatrix} L_1 & & \\ L_2 & I & \\ & & I \end{bmatrix} \begin{bmatrix} L_1^{-1}G_{11}L_1^{-T} & L_1^{-1}G_1Z & U \\ Z^TG_1^TL_1^{-T} & Z^TGZ & \\ U^T & & \end{bmatrix} \begin{bmatrix} L_1^T & L_2^T & \\ & I & \\ & & I \end{bmatrix},$$

where Z is defined by (3.2) and $G_1 = [G_{11}\ G_{12}]$. Note that the matrix U occurs on the reverse diagonal of the middle factor but that no operations with U^{-1} are required in the calculation of the factors. Thus no ill conditioning associated with U manifests itself until the factors are used in solving the KKT system (1.1). If there is no growth in Z, then the backward error in (3.7) will be small, indicating the potential for a small residual solution of the KKT system. We show in section 5 how this can come about. Another related observation is that if A is rank deficient, then the factors (3.6) do not exist (since the calculation of Y involves A_1^{-1} and hence U^{-1}), whereas (3.7) can be calculated without difficulty.

The factorization (3.7) of K is closely related to some symmetry preserving variants of Gaussian elimination. Let us start by eliminating A_2 and the subdiagonal

elements of A_1 by row operations. (As before, we can assume that row pivoting has been used.) The outcome of these row operations is that

$$\begin{bmatrix} G_{11} & G_{12} & A_1 \\ G_{21} & G_{22} & A_2 \\ A_1^T & A_2^T & 0 \end{bmatrix} = \begin{bmatrix} L_1 & & \\ L_2 & I & \\ & & I \end{bmatrix} \begin{bmatrix} L_1^{-1}G_{11} & L_1^{-1}G_{12} & U \\ Z^T G_1^T & Z^T G_2^T & \\ A_1^T & A_2^T & \end{bmatrix}, \tag{3.8}$$

where $G_2 = [G_{21}\ G_{22}]$. Note that these row operations are exactly those used by Gaussian elimination to form (3.1). To restore symmetry in the factors, we repeat the above procedure in transposed form; that is, we make column operations on A_1^T and A_2^T, which gives rise to (3.7).

We can also interleave these row and column operations without affecting the final result. If we pair up the first row and column operation, then the second row and column operation, and so on, then we get the method of "HA" pivots described by Forsgren and Murray [7]. Thus these methods essentially share the same matrix factors. The difference is that in the null-space method, $Z^T G Z$ is calculated by matrix solves with $\mathbf{A}$, as described in section 2, whereas in these other methods it is obtained by row and column operations on the matrix K. This result of this paragraph is also observed by Gill et al. [8].

This association with Gaussian elimination enables us to bound the growth in the factors of K. The bound is attained for the critical case typified by the matrix

$$K = \left[\begin{array}{cccccc|cccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & -1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 0 & 1 & -1 & -1 & -1 & -1 \\ \hline 1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \end{array}\right],$$

for which $n = 6$ and $m = 4$. Row operations with pivots in the (1,7), (2,8), (3,9), and (4,10) positions lead to the matrix

$$\left[\begin{array}{cccccc|cccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 2 & 2 & 2 & 2 & 2 & 2 & 0 & 1 & 0 & 0 \\ 4 & 4 & 4 & 4 & 4 & 4 & 0 & 0 & 1 & 0 \\ 8 & 8 & 8 & 8 & 8 & 8 & 0 & 0 & 0 & 1 \\ 16 & 16 & 16 & 16 & 16 & 15 & 0 & 0 & 0 & 0 \\ 16 & 16 & 16 & 16 & 15 & 16 & 0 & 0 & 0 & 0 \\ \hline 1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \end{array}\right].$$

Then column operations with pivots in the (7,1), (8,2), (9,3), and (10,4) positions

give rise to

$$\left[\begin{array}{cccccc|cccc} 1 & 2 & 4 & 8 & 16 & 16 & 1 & 0 & 0 & 0 \\ 2 & 4 & 8 & 16 & 32 & 32 & 0 & 1 & 0 & 0 \\ 4 & 8 & 16 & 32 & 64 & 64 & 0 & 0 & 1 & 0 \\ 8 & 16 & 32 & 64 & 128 & 128 & 0 & 0 & 0 & 1 \\ 16 & 32 & 64 & 128 & 256 & 255 & 0 & 0 & 0 & 0 \\ 16 & 32 & 64 & 128 & 255 & 256 & 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right],$$

which corresponds to the middle factor in (3.7). In this case $U = I$, $L = A$, and the corresponding matrix Z is given by

$$Z^T = \begin{bmatrix} 8 & 4 & 2 & 1 & 1 & 0 \\ 8 & 4 & 2 & 1 & 0 & 1 \end{bmatrix}.$$

In general it is readily shown that when $m < n$, growth of 2^{2m} in the maximum modulus element of K can occur. Of course, if K were factorized using some pivots from G, this growth would no longer arise, but this would not be a null-space method. For the null-space method based on (3.2), this example also illustrates the maximum possible growth of 2^{m-1} in Z, when $|l_{ij}| \leq 1$. In practice, however, such growth is most unlikely and it is usual not to get any significant growth in Z.

4. Numerical stability of Method 1. In this and the next section we consider the effect of ill conditioning in the matrix K on the solutions obtained by null-space methods based on direct elimination. In particular, we are interested to see whether or not we can establish results comparable to those for Gaussian elimination. We shall show that the forward error in $\mathbf{x}$ is not as severe as would be predicted by the condition number of K. We also look at the residual errors in the solution and show that Method 2 is very satisfactory in this respect, whereas Method 1 is not.

In order to prevent the details of the analysis from obscuring the insight that we are trying to provide, we shall adopt the following simple convention. We imagine that we are solving a sequence of problems in which either $\kappa_{\mathbf{A}}$ or κ_M (the spectral condition numbers of $\mathbf{A}$ and $M = Z^T G Z$) is increasing without bound. We then use the notation $O(h)$ to indicate a quantity whose norm is bounded by $c\|h\|$ on this sequence, where there exists an implied constant c that is independent of $\kappa_{\mathbf{A}}$ or κ_M, but may contain a modest dependence on n. Also, we shall assume that the system is well scaled so that $G = O(1)$ and $\mathbf{A} \sim 1$ (defined by $\mathbf{A} = O(1)$ and $\|\mathbf{A}\|^{-1} = O(1)$). This enables us, for example, to deduce that multiplication of an error bound $O(\varepsilon)$ by $\mathbf{A}^{-1}$ causes the bound to be increased to $O(\kappa_{\mathbf{A}}\varepsilon)$. We also choose to assume that the KKT system models a situation in which the exact solution vectors $\mathbf{x}$ and $\mathbf{y}$ exist and are not unreasonably large in norm; that is, $\mathbf{x} = O(1)$ and $\mathbf{y} = O(1)$. A similar assumption is needed in order to show that Gaussian elimination provides accurate residuals, so we cannot expect to dispense with this assumption. Sometimes it may be possible to argue that we are solving a physical problem which is known to have a well-behaved solution. Also, in QP applications it is usual for the iterates $\mathbf{x}$ to be of reasonable magnitude.

Another assumption that we make is that the choice of the matrix V in (2.8) (and hence the partitioning of A) is made using some form of pivoting. Now the exact solution for Z is given by

$$Z^T = [\,-A_2A_1^{-1} \quad I\,] = [\,-L_2L_1^{-1} \quad I\,]$$

from (3.3), using the factors of A defined in (3.1). It follows that

$$Z = O(\kappa_{\mathbf{L}}), \tag{4.1}$$

where $\kappa_{\mathbf{L}}$ is the spectral condition number of $\mathbf{L}$. Assuming that partial pivoting is used, so that $|l_{ij}| \leq 1$, and that negligible growth occurs in L_1^{-1}, it then follows that negligible growth occurs in Z, and we can assert that

$$\kappa_{\mathbf{L}} = O(1) \quad \text{and} \quad Z = O(1). \tag{4.2}$$

Another consequence of this assumption is that we are able to neglect terms of $O(\kappa_{\mathbf{L}}\varepsilon)$ relative to terms of $O(\kappa_{\mathbf{A}}\varepsilon)$ when assessing the propagation of errors for Method 2.

We shall now sketch some properties of floating point arithmetic of relative precision ε. If a nonsingular system of n linear equations $A\mathbf{x} = \mathbf{b}$ is solved by Gaussian elimination, the computed solution $\widehat{\mathbf{x}}$ is the exact solution of a perturbed system $(A + E)\widehat{\mathbf{x}} = \mathbf{b}$, where E is referred to as the backward error (Wilkinson [14, section 4.29]). E can be bounded by an expression of the form $\rho\phi(n)\varepsilon + O(\varepsilon^2)$ in which ρ measures the growth in A during the elimination and $\phi(n)$ is a modest quadratic in n (Stewart [11, Theorem 5.3]). This bound usually overstates the dependence on n which is unlikely to be a dominant factor. Also for ill-conditioned systems, and assuming that partial pivoting is used, it is usual for the size of the reduced matrices to diminish (Wilkinson [14, sections 4.40 and 4.31]). Thus significant growth is rare and to simplify the presentation of our results we assume that it does not occur. Hence, for the backward error, we may write

$$E = O(\varepsilon). \tag{4.3}$$

We can measure the accuracy of the solution either by the forward error $\widehat{\mathbf{x}} - \mathbf{x} = -A^{-1}E\widehat{\mathbf{x}}$ or by computing the residual $\mathbf{r} = A\widehat{\mathbf{x}} - \mathbf{b} = -E\widehat{\mathbf{x}}$. Using $\mathbf{A} \sim 1$ we have

$$\widehat{\mathbf{x}} = \mathbf{x} + O(\kappa_A\varepsilon\widehat{\mathbf{x}}),$$

where κ_A is some condition number of A. Since $\mathbf{x} = O(1)$, and assuming that $\kappa_A\varepsilon \ll 1$, it follows that $\widehat{\mathbf{x}} = O(1)$ and hence

$$\widehat{\mathbf{x}} = \mathbf{x} + O(\kappa_A\varepsilon). \tag{4.4}$$

Likewise, we can deduce that

$$\mathbf{r} = -E\widehat{\mathbf{x}} = O(\varepsilon). \tag{4.5}$$

These bounds are likely to be realistic and tell us that for Gaussian elimination, ill conditioning affects the forward error in $\mathbf{x}$ but not the residual $\mathbf{r}$, as long as $\widehat{\mathbf{x}}$ is of reasonable magnitude.

Wilkinson [13, section 1.26] gives expressions for the backward error in a scalar product and hence in the product $\mathbf{s} = \mathbf{b} + A\mathbf{x}$. The computed product $\widehat{\mathbf{s}}$ is the exact

product of a system in which the relative perturbation in each element of $\mathbf{b}$ and A is no more than $n\varepsilon$ where n is the dimension of $\mathbf{x}$. We can express this as

$$\widehat{\mathbf{s}} = \mathbf{s} + O(\varepsilon) \tag{4.6}$$

if we make the assumption that $\mathbf{b}$ and A are $O(1)$.

The first stage in a null-space calculation is the determination of Z^TGZ, which we denote by M. In Method 1, this is computed as in (2.10) and (2.11). In (2.10) a column $\mathbf{z}_k$ of the matrix Z is computed which, by applying (4.4), satisfies

$$\widehat{\mathbf{z}}_k = \mathbf{z}_k + O(\kappa_{\mathbf{A}}\varepsilon), \tag{4.7}$$

where $\kappa_{\mathbf{A}}$ is the spectral condition number of $\mathbf{A}$. The product with G introduces negligible error, and the solution of (2.11) together with (4.5) shows that

$$\mathbf{A}\widehat{\mathbf{u}} = G\widehat{\mathbf{z}}_k + O(\varepsilon).$$

Multiplying by $\mathbf{L}^{-1}$ and extracting the $\widehat{\mathbf{u}}_2$ partition gives

$$\begin{aligned}\widehat{\mathbf{u}}_2 &= Z^TG\widehat{\mathbf{z}}_k + O(\kappa_{\mathbf{L}}\varepsilon)\\ &= Z^TG\mathbf{z}_k + O(\kappa_{\mathbf{A}}\varepsilon)\end{aligned}$$

using (4.7) and then (4.2). Hence we have established that

$$\widehat{M} = M + O(\kappa_{\mathbf{A}}\varepsilon). \tag{4.8}$$

The argument has been given in some detail as it is important to see why the error in M is $O(\kappa_{\mathbf{A}}\varepsilon)$ and not $O(\kappa_{\mathbf{A}}^2\varepsilon)$. We also observe that $M = Z^TGZ = O(1)$ and hence that $\widehat{M} = O(1)$ when $\kappa_{\mathbf{A}}\varepsilon \ll 1$.

We now turn to the solution of the KKT system using Method 1. We shall assume that systems involving $\mathbf{A}$ and M are solved in such a way that (4.5) applies. Using (4.6) and assuming that the computed quantities $\widehat{\mathbf{s}}, \widehat{\mathbf{t}}, \ldots, \widehat{\mathbf{z}}$ are $O(1)$, the residual errors in the sequence of calculations are then

$$\mathbf{A}^T\widehat{\mathbf{s}} = \begin{pmatrix}\mathbf{b}\\ \mathbf{0}\end{pmatrix} + O(\varepsilon), \tag{4.9}$$

$$\widehat{\mathbf{t}} = \mathbf{c} - G\widehat{\mathbf{s}} + O(\varepsilon), \tag{4.10}$$

$$\mathbf{A}\widehat{\mathbf{u}} = \widehat{\mathbf{t}} + O(\varepsilon), \tag{4.11}$$

$$\widehat{M}\widehat{\mathbf{v}} = \widehat{\mathbf{u}}_2 + O(\varepsilon), \tag{4.12}$$

$$\mathbf{A}^T\widehat{\mathbf{x}} = \begin{pmatrix}\mathbf{b}\\ \widehat{\mathbf{v}}\end{pmatrix} + O(\varepsilon), \tag{4.13}$$

$$\widehat{\mathbf{g}} = \mathbf{c} - G\widehat{\mathbf{x}} + O(\varepsilon), \tag{4.14}$$

$$\mathbf{A}\begin{pmatrix}\widehat{\mathbf{y}}\\ \widehat{\mathbf{z}}\end{pmatrix} = \widehat{\mathbf{g}} + O(\varepsilon). \tag{4.15}$$

These results, together with (4.8), may be combined to get the forward errors in the solution vectors $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{y}}$. Multiplying through equations (4.9) and (4.13) by $\mathbf{A}^{-T}$ magnifies the error bounds by a factor $\kappa_{\mathbf{A}}$ (since we are assuming that $\mathbf{A} \sim 1$), giving

$$\widehat{\mathbf{s}} = Y\mathbf{b} + O(\kappa_{\mathbf{A}}\varepsilon), \tag{4.16}$$

$$\widehat{\mathbf{x}} = Y\mathbf{b} + Z\widehat{\mathbf{v}} + O(\kappa_{\mathbf{A}}\varepsilon). \tag{4.17}$$

We can get a rather better bound from (4.11) and (4.15) by first multiplying through by $\mathbf{L}^{-1}$ and using $\kappa_{\mathbf{L}} = O(1)$ to give

$$\widehat{\mathbf{u}}_2 = Z^T\widehat{\mathbf{t}} + O(\varepsilon), \tag{4.18}$$

$$\widehat{\mathbf{z}}_2 = Z^T\widehat{\mathbf{g}} + O(\varepsilon) \tag{4.19}$$

from the second partition of the solution. However, the first partition of (4.15) gives

$$\widehat{\mathbf{y}} = Y^T\widehat{\mathbf{g}} + O(\kappa_{\mathbf{A}}\varepsilon). \tag{4.20}$$

Combining (4.8) and (4.12) gives

$$\widehat{\mathbf{v}} = M^{-1}\widehat{\mathbf{u}}_2 + O(\kappa_{\mathbf{A}}\varepsilon) + O(\kappa_M\varepsilon). \tag{4.21}$$

We can now chain through the forward errors, noting that a product with Z or Z^T does not magnify the order of the error bound in a previously computed quantity (by virtue of (4.2)). However, the product $M^{-1}\widehat{\mathbf{u}}_2$ in (4.21) magnifies the error bound in $\widehat{\mathbf{u}}_2$ by a factor κ_M, and the product $Y^T\widehat{\mathbf{g}}$ in (4.20) magnifies the error bound in $\widehat{\mathbf{g}}$ by a factor $\kappa_{\mathbf{A}}$. The outcome is that

$$\widehat{\mathbf{x}} = \mathbf{x} + O(\kappa_{\mathbf{A}}\kappa_M\varepsilon) \tag{4.22}$$

and

$$\widehat{\mathbf{y}} = \mathbf{y} + O(\kappa_{\mathbf{A}}^2\kappa_M\varepsilon). \tag{4.23}$$

As we would expect, the forward errors are affected by the condition numbers of $\mathbf{A}$ and M. However, although the condition number of K is expected to be of the order $\kappa_{\mathbf{A}}^2\kappa_M$, we see that this factor only magnifies the error bound in the $\mathbf{y}$ part of the solution, with the $\mathbf{x}$ part being less badly affected.

When K is ill conditioned we must necessarily expect that the forward errors are adversely affected. A more important question is to ask whether the solution satisfies the equations of the problem accurately. There are three measures of interest, the residuals $\mathbf{q} = G\mathbf{x} + A\mathbf{y} - \mathbf{c}$ and $\mathbf{r} = A^T\mathbf{x} - \mathbf{b}$ of the KKT system (1.1) and the reduced gradient $\mathbf{z} = Z^T\mathbf{g}$ where $\mathbf{g} = \mathbf{c} - G\mathbf{x}$ is the negative gradient vector at the solution. We note that the vector $\mathbf{z}$ is computed as a by-product of step 8 of Method 1.

If we compute $\mathbf{r}$ we obtain $\widehat{\mathbf{r}} = A^T\widehat{\mathbf{x}} - \mathbf{b} + O(\varepsilon)$ as in (4.6), and it follows from (4.13) and the definition of $\mathbf{A}$ that $A^T\widehat{\mathbf{x}} = \mathbf{b} + O(\varepsilon)$. Thus

$$\widehat{\mathbf{r}} = O(\varepsilon). \tag{4.24}$$

When computing $\mathbf{q}$ we obtain

$$\widehat{\mathbf{q}} = G\widehat{\mathbf{x}} + A\widehat{\mathbf{y}} - \mathbf{c} + O(\varepsilon) \tag{4.25}$$

$$= A\widehat{\mathbf{y}} - \widehat{\mathbf{g}} + O(\varepsilon) \tag{4.26}$$

$$= -\begin{pmatrix} \mathbf{0} \\ \widehat{\mathbf{z}} \end{pmatrix} + O(\varepsilon) \tag{4.27}$$

from (4.14) and (4.15). Thus the accuracy of $\widehat{\mathbf{q}}$ depends on that of $\widehat{\mathbf{z}}$. From (4.19) and (4.14) it follows that

$$\begin{aligned} \widehat{\mathbf{z}} &= Z^T\mathbf{c} - Z^TG\widehat{\mathbf{x}} + O(\varepsilon) \\ &= Z^T\mathbf{c} - Z^TGY\mathbf{b} - Z^TGZ\widehat{\mathbf{v}} + O(\kappa_{\mathbf{A}}\varepsilon) \end{aligned}$$

from (4.17). (Notice that it is important not to use (4.22) here, which would give an unnecessary factor of κ_M.) Then (4.8), (4.12), (4.11), and (4.10) can be used, giving

$$\widehat{\mathbf{z}} = O(\kappa_{\mathbf{A}}\varepsilon). \tag{4.28}$$

Thus we are able to predict under our assumptions that the reduced gradient $\widehat{\mathbf{z}}$ and the residual $\widehat{\mathbf{q}}$ are adversely affected by ill conditioning in A but not by ill conditioning in M. However, the residual $\widehat{\mathbf{r}}$ is unaffected by ill conditioning either in A or M.

Simulations are described in section 6 which indicate that these error bounds reliably predict the actual effects of ill conditioning. Method 1 is seen to be unsatisfactory in that an accurate residual $\mathbf{q}$ cannot be obtained when A is ill conditioned. We shall show in the next section that Method 2 does not share this disadvantage.

The main results of this section and the next are summarized and discussed in section 7.

5. Numerical stability of Method 2. In this section we assess the behavior of Method 2 in the presence of ill conditioning in K. Although we cannot expect any improvement for the forward errors, we are able to show that Method 2 is able to give accurate residuals that are not affected by ill conditioning. The relationship between Method 2 and Gaussian elimination described towards the end of section 3 gives some hope of proving this result. However, this is not immediate because Method 2 does not make direct use of the factors (3.7) in the same way that Gaussian elimination does.

A fundamental difficulty with the analysis of Method 2 is that we can deduce from (4.7) that

$$\widehat{Z} = Z + O(\kappa_{\mathbf{A}}\varepsilon), \tag{5.1}$$

and this result cannot be improved if LU factors are available. To see this, we know that the computed factors of any square matrix A satisfy

$$\widehat{L}\widehat{U} = A + E = A + O(\varepsilon) \tag{5.2}$$

when there is no growth in $\widehat{U}$. If $A = LU$ are the exact factors, it follows that

$$L^{-1}\widehat{L} = U\widehat{U}^{-1} + L^{-1}E\widehat{U}^{-1} = U\widehat{U}^{-1} + Q + R,$$

say, where Q is the strict lower triangular part of $L^{-1}E\widehat{U}^{-1}$ and R is the upper triangular part. Because $L^{-1}\widehat{L}$ is unit lower triangular and $U\widehat{U}^{-1}$ is upper triangular we can deduce that $L^{-1}\widehat{L} = I + Q$ and $U\widehat{U}^{-1} = I - R$. Since $L^{-1}E\widehat{U}^{-1}$ involves an inverse operation with $\widehat{U}$ we can expect that $\widehat{L}$ and L differ by $O(\kappa_{\mathbf{A}}\varepsilon)$. This result has been confirmed by computing the LU factors of a Hilbert matrix in single and double precision Fortran. On applying the result to our matrix $\mathbf{A}$, it follows that (5.1) holds.

Fortunately all is not lost because we are still able to compute a null-space matrix which accurately satisfies the equation $Z^TA = 0$. Let $\widetilde{Z}$ denote the null-space matrix obtained from $\widehat{L}$ in exact arithmetic. It follows that $\widetilde{Z}^T\widehat{L} = 0$ and hence from (5.2) that

$$\widetilde{Z}^TA = O(\varepsilon). \tag{5.3}$$

We also have $\widetilde{Z} = O(1)$ as long as $\kappa_{\mathbf{A}}\varepsilon \ll 1$. Because we use the matrix $\widehat{L}$ in computing the solutions of (5.7), (5.9), and (5.13) below, we can use $\widetilde{Z}$ rather than Z in the analysis, and this enables us to avoid the $\kappa_{\mathbf{A}}$ factor in the residuals.

The first step in Method 2 is to compute $M = Z^T G Z$ as in (3.4) and (3.5). In this section we denote $\widetilde{M} = \widetilde{Z}^T G \widetilde{Z}$ as the value computed from $\widetilde{Z}$ in exact arithmetic and retain $\widehat{M}$ to denote the computed value of M. It readily follows, using results like (4.2), that

$$\widehat{M} = \widetilde{M} + O(\varepsilon). \tag{5.4}$$

We now consider the solution of the KKT system using Method 2. As in equations (4.9) through (4.15) we assume that the computed quantities $\widehat{\mathbf{s}}$, $\widehat{\mathbf{t}}, \ldots, \widehat{\mathbf{z}}$ are $O(1)$. Then the residual errors in the sequence of calculations are

$$A_1^T \widehat{\mathbf{s}}_1 = \mathbf{b} + O(\varepsilon) \quad \text{and} \quad \widehat{\mathbf{s}}_2 = \mathbf{0}, \tag{5.5}$$

$$\widehat{\mathbf{t}} = \mathbf{c} - G\widehat{\mathbf{s}} + O(\varepsilon), \tag{5.6}$$

$$\widehat{\mathbf{u}}_2 = \widetilde{Z}^T \widehat{\mathbf{t}} + O(\varepsilon), \tag{5.7}$$

$$\widehat{M}\widehat{\mathbf{v}} = \widehat{\mathbf{u}}_2 + O(\varepsilon), \tag{5.8}$$

$$\widehat{\mathbf{w}} = \widetilde{Z}\widehat{\mathbf{v}} + O(\varepsilon), \tag{5.9}$$

$$\widehat{\mathbf{x}} = \widehat{\mathbf{s}} + \widehat{\mathbf{w}} + O(\varepsilon), \tag{5.10}$$

$$\widehat{\mathbf{g}} = \mathbf{c} - G\widehat{\mathbf{x}} + O(\varepsilon), \tag{5.11}$$

$$A_1 \widehat{\mathbf{y}} = \widehat{\mathbf{g}}_1 + O(\varepsilon), \tag{5.12}$$

$$\widehat{\mathbf{z}} = \widetilde{Z}^T \widehat{\mathbf{g}} + O(\varepsilon). \tag{5.13}$$

It is readily seen from these equations that the forward errors will propagate in a similar way to Method 1.

Turning to the residual errors, the computed value of the residual $\mathbf{r}$ is

$$\widehat{\mathbf{r}} = A^T \widehat{\mathbf{x}} - \mathbf{b} + O(\varepsilon) = A^T \widehat{\mathbf{s}} + A^T \widetilde{Z}\widehat{\mathbf{v}} - \mathbf{b} + O(\varepsilon) = O(\varepsilon) \tag{5.14}$$

from (5.10), (5.9), (5.5), and (5.3). When computing $\mathbf{q}$ we obtain $\widehat{\mathbf{q}} = A\widehat{\mathbf{y}} - \widehat{\mathbf{g}} + O(\varepsilon)$ as for Method 1, and it follows from (5.12) that $\widehat{\mathbf{q}}_1 = O(\varepsilon)$. From (5.3) we can deduce that $\widetilde{Z}^T \widehat{\mathbf{q}} = -\widetilde{Z}^T \widehat{\mathbf{g}} + O(\varepsilon)$. But $\widetilde{Z}^T \widehat{\mathbf{q}} = \widehat{\mathbf{q}}_2 - \widehat{L}_1^{-1} \widehat{L}_2 \widehat{\mathbf{q}}_1 = \widehat{\mathbf{q}}_2 + O(\varepsilon)$, so it follows that

$$\widehat{\mathbf{q}}_2 = -\widetilde{Z}^T \widehat{\mathbf{g}} + O(\varepsilon) = -\widehat{\mathbf{z}} + O(\varepsilon). \tag{5.15}$$

Thus the accuracy of the residual $\widehat{\mathbf{q}}$ depends on that of $\widehat{\mathbf{z}}$, as for Method 1. For $\widehat{\mathbf{z}}$ we can use (5.13), (5.11), (5.10), and (5.9) to get

$$\widehat{\mathbf{z}} = \widetilde{Z}^T \mathbf{c} - \widetilde{Z}^T G \widehat{\mathbf{s}} - \widetilde{Z}^T G \widetilde{Z} \widehat{\mathbf{v}} + O(\varepsilon).$$

Now we can invoke (5.4) and (5.8), giving

$$\widehat{\mathbf{z}} = \widetilde{Z}^T \mathbf{c} - \widetilde{Z}^T G \widehat{\mathbf{s}} - \widehat{\mathbf{u}}_2 + O(\varepsilon) = O(\varepsilon) \tag{5.16}$$

from (5.7) and (5.6). Thus we have established under our assumptions that all three measures of accuracy for the KKT system are $O(\varepsilon)$ for Method 2 and are not affected by ill conditioning in either A or M. These results are again supported by the simulations in the next section.

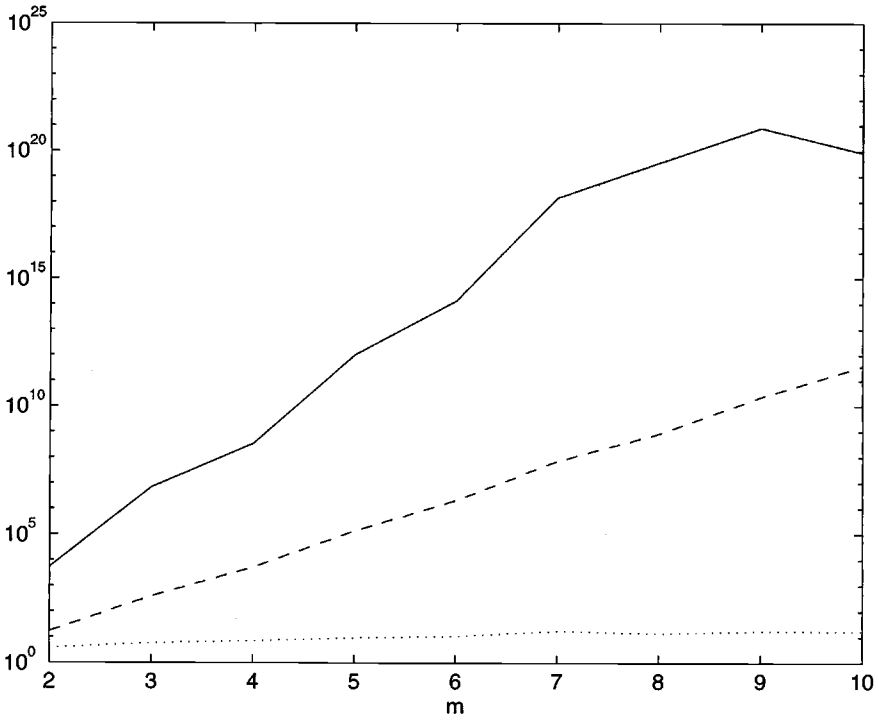


FIG. 1. *Condition numbers of* K, $\mathbf{A}$, *and* $\mathbf{L}$.

6. Numerical experiments. In order to check the predictions of sections 4 and 5, some experiments have been carried out on artificially generated KKT systems. These experiments have been carried out in Matlab, for which the machine precision is $\varepsilon \simeq 10^{-16}$. They suggest that the upper bounds given by the error analysis accurately reflect the actual behavior of an ill-conditioned system. Another phenomenon that occurs when the ill conditioning is very extreme is also explained.

The KKT systems have been constructed in the following way. To make A ill conditioned we have chosen it as the first m columns of the $n \times n$ Hilbert matrix, where $n = 2m$. Choosing $m = 2, 3, \ldots, 10$ provides a sequence of problems for which the condition number of $\mathbf{A}$ increases exponentially. Factors $PA = LU$ are calculated by the Matlab routine `lu`, which uses Gaussian elimination with partial pivoting, and A is replaced by PA. In the first instance the matrix G is generated by random numbers in the range $[-1, 1]$. However, to make $M = Z^T G Z$ positive definite, a multiple of the unit matrix is added to the G_{22} partition of G, chosen so that the smallest eigenvalue of M is changed to 10^{1-k} for some positive integer k. The assumptions of the analysis require that the KKT system has a solution that is $O(1)$. To achieve this, exact solutions $\mathbf{x}$ and $\mathbf{y}$ are generated by random numbers in $[-1, 1]$, and the right-hand sides $\mathbf{c}$ and $\mathbf{b}$ are calculated from (1.1). For each value of m, 10 runs are made with a different random number seed and the statistics are averaged over these 10 runs.

First, we examine the effect of increasing the condition number of A while keeping M well conditioned. To do this we increase m from 2 up to 10, while fixing $k = 1$. The resulting condition numbers of K, $\mathbf{A}$, and $\mathbf{L}$ are plotted in Figure 1. It can be seen that the slope of the unbroken line (κ_K) is about twice that of the dashed line ($\kappa_{\mathbf{A}}$). Since $\kappa_M \sim 1$, this is consistent with the estimate $\kappa_K \sim \kappa_{\mathbf{A}}^2 \kappa_M$ that we deduced in

section 3. The condition number of $\mathbf{L}$ (dotted line) shows negligible increase, showing that there is no growth in L_1^{-1}, thus enabling us to assert that $Z = O(1)$. The leveling out of the κ_K graph for $m = 8$, 9, and 10 is due to round-off error corrupting the least eigenvalue of K.

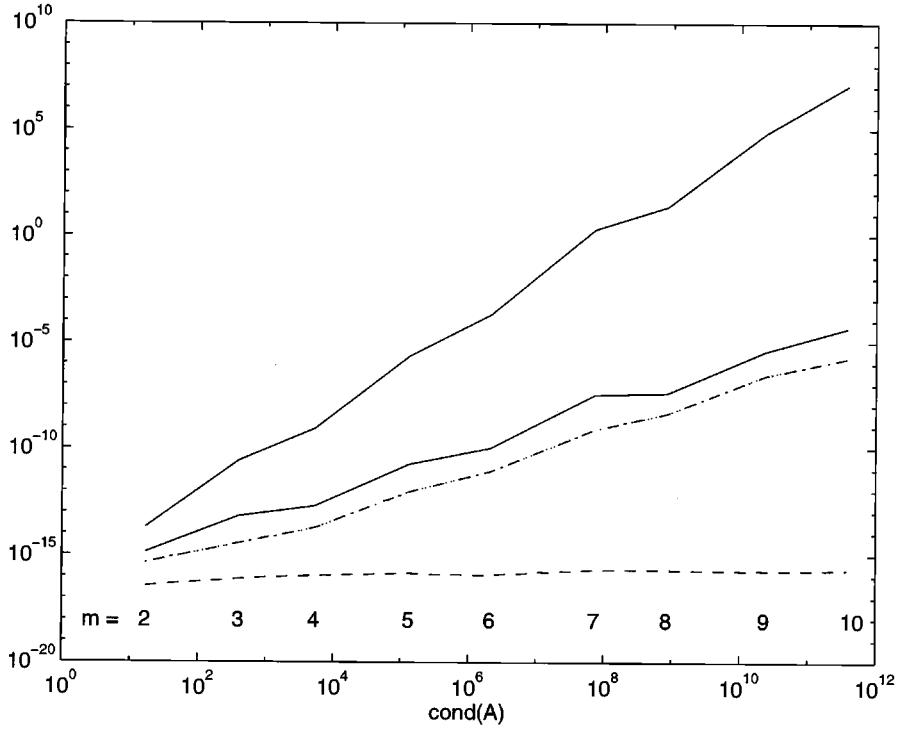


FIG. 2. *Error growth vs.* $\kappa_{\mathbf{A}}$ *for Method* 1.

The effect of the conditioning of A on the different types of error is illustrated in Figures 2 and 3. The forward error is shown by the two unbroken lines, the upper line being the error in $\mathbf{y}$ and the lower line being the error in $\mathbf{x}$. The upper line has a slope of about 2 on the log-log scale, and the lower line has a slope of about 1, and both have an intercept with the y-axis of about 10^{-16}. This is precisely in accordance with (4.23) and (4.22). It can also be seen that both methods exhibit the same pattern of behavior in the forward error. The computed value of the residual error $\mathbf{r} = A^T\mathbf{x} - \mathbf{b}$ is shown by the dashed line and both methods show the $O(\varepsilon)$ behavior as predicted by (4.24) and (5.14), with the increasing condition number having no effect.

The difference between Methods 1 and 2 is shown by the computed values of the residual $\mathbf{q} = G\mathbf{x} + A\mathbf{y} - \mathbf{c}$ (dotted line) and the reduced gradient $\mathbf{z} = Z^T\mathbf{g}$ (dash-dot line). As we would expect from (4.27), these graphs are superimposed, and they clearly show the influence of $\kappa_{\mathbf{A}}$ on the error growth for Method 1, as predicted by (4.28). Negligible error growth is observed for Method 2 as predicted by (5.16), except for an increase in $\mathbf{q}$ for $\kappa_{\mathbf{A}}$ greater than about 10^9. This feature is explained later in the section.

We now turn to see the influence of ill conditioning in M on the errors. To do this we fix $m = 5$, for which $\kappa_{\mathbf{A}} \simeq 10^5$, and carry out a sequence of calculations with $k = 1, 2, \ldots, 10$, which causes κ_M to increase exponentially. Each calculation is the

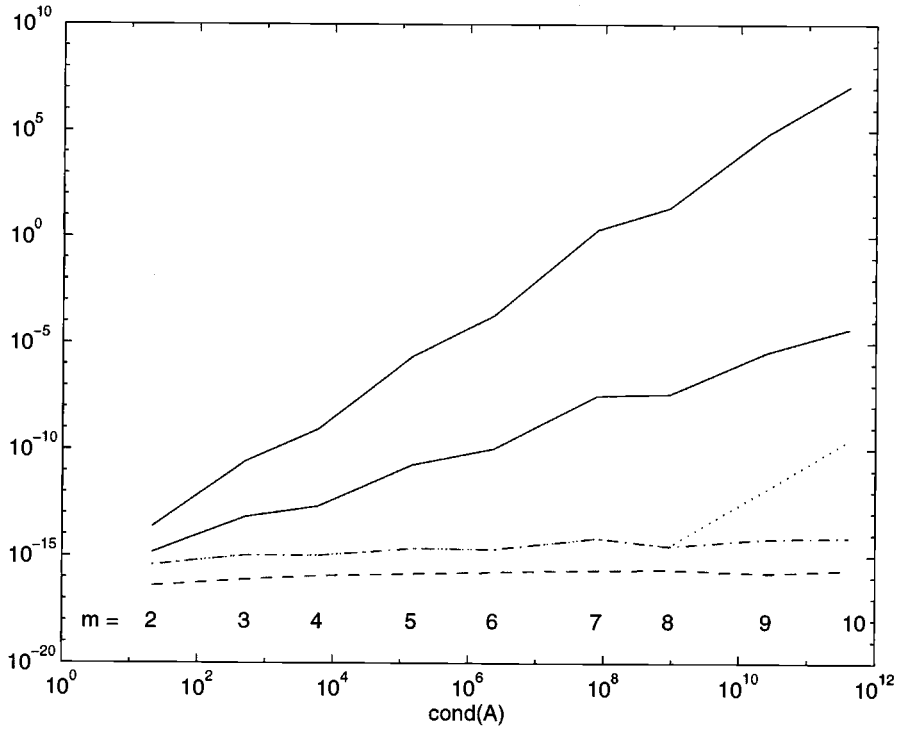


FIG. 3. *Error growth vs.* $\kappa_{\mathbf{A}}$ *for Method* 2.

average of 10 runs, as above. The results are illustrated in Figures 4 and 5, using the same key. The forward errors are again seen to be similar for both methods and they both have a slope of about 1 on the log-log scale, corresponding to the κ_M factor in (4.22) and (4.23). The upper line for the forward error in $\mathbf{y}$ lies about 10^5 units above that for the forward error in $\mathbf{x}$, as the extra factor of $\kappa_{\mathbf{A}}$ in (4.23) would predict. The residual $\mathbf{r}$ is seen to be unaffected by the conditioning of M as above. The residual $\mathbf{q}$ and the reduced gradient $\mathbf{z}$ are also unaffected by κ_M, but the graphs for Method 1 lie above those for Method 2, due to the $\kappa_{\mathbf{A}}$ factor in (4.28). All these effects are in accordance with what the error analysis predicts.

To examine the anomalous behavior of $\mathbf{q}$ in Figure 3 in more detail, we turn to a sequence of more ill-conditioned test problems obtained by using the *last* m columns of the Hilbert matrix to define A. The results for Method 2 are illustrated in Figure 6, and the anomalous behavior (dotted line) is now very evident. The reason for this becomes apparent when it is noticed that it sets in when the forward error in $\mathbf{y}$, and hence the value of $\widehat{\mathbf{y}}$, becomes greater than unity. This possibility has been excluded in our error analysis by the assumption that $\widehat{\mathbf{y}} = O(1)$. The anomalous behavior sets in when $\kappa_{\mathbf{A}}^2 \kappa_M \varepsilon \simeq 1$, that is, $\kappa_{\mathbf{A}} \simeq (\kappa_M \varepsilon)^{-1/2}$, or in this case $\kappa_{\mathbf{A}} \simeq 10^8$, much as Figures 3 and 6 illustrate. For greater values of $\kappa_{\mathbf{A}}$ there is a term $O(\widehat{\mathbf{y}}\varepsilon)$ in the expression for $\widehat{\mathbf{q}}$ indicating that the error is of the form $\kappa_{\mathbf{A}}^2 \kappa_M \varepsilon^2$. The fact that this part of the graph of $\widehat{\mathbf{q}}$ is parallel to the graph of the forward error in $\mathbf{y}$ supports this conclusion.

The above calculations have also been carried out using a Vandermonde matrix in place of the Hilbert matrix, and very similar results have been obtained.

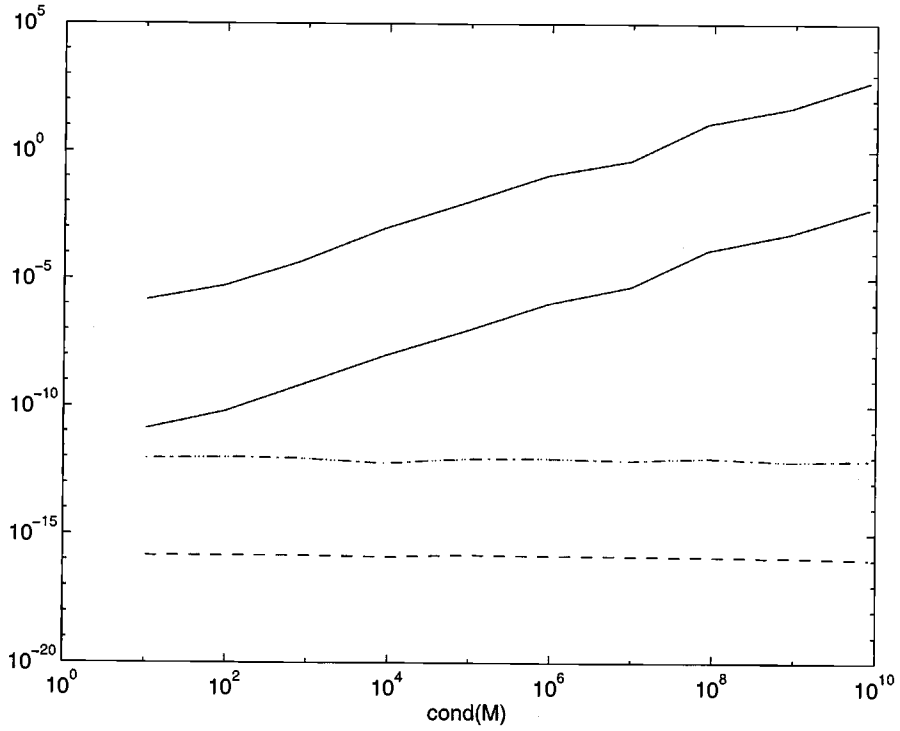


FIG. 4. *Error growth vs.* κ_M *for Method* 1.

In this paper we have examined the effect of ill conditioning on the solution of a KKT system by null-space methods based on direct elimination. Such methods are important because they are well suited to take advantage of sparsity in large systems. However, they have often been criticized for a lack of numerical stability, particularly when compared to methods based on QR factors. We have studied two methods: Method 1, in which an invertible representation of $\mathbf{A}$ in (2.8) is used to solve systems, and Method 2, in which LU factors (3.1) of A are available. We have presented error analysis backed up by numerical simulations which, under certain assumptions on growth and the size of the solutions, provide the following conclusions.

- Both methods have the same forward error bounds, with $\widehat{\mathbf{x}} = \mathbf{x} + O(\kappa_{\mathbf{A}}\kappa_M\varepsilon)$ and $\widehat{\mathbf{y}} = \mathbf{y} + O(\kappa_{\mathbf{A}}^2\kappa_M\varepsilon)$.
- Both methods give accurate residuals if A is well conditioned, even if M is ill conditioned.
- Method 2 always gives an accurate residual $\mathbf{q} = G\mathbf{x} + A\mathbf{y} - \mathbf{c}$, whereas $\mathbf{q} = O(\kappa_{\mathbf{A}}\varepsilon)$ for Method 1.
- Both methods give an accurate residual $\mathbf{r} = A^T\mathbf{x} - \mathbf{b}$ if A is ill conditioned.

These conclusions do indicate that Method 1 is adversely affected by ill conditioning in A, even though the technique for solving systems involving $\mathbf{A}$ is able to provide accurate residuals. The reasons for this are particularly interesting. For example, one might expect that when A is ill conditioned, then $\mathbf{A}^{-1}$ would be large and we might, therefore, expect from (2.1) that Z would be large. In fact, we have seen that as long as V is chosen suitably, then growth in Z is very unlikely (the argument is similar to that for Gaussian elimination). Of course, if V is badly chosen then Z can

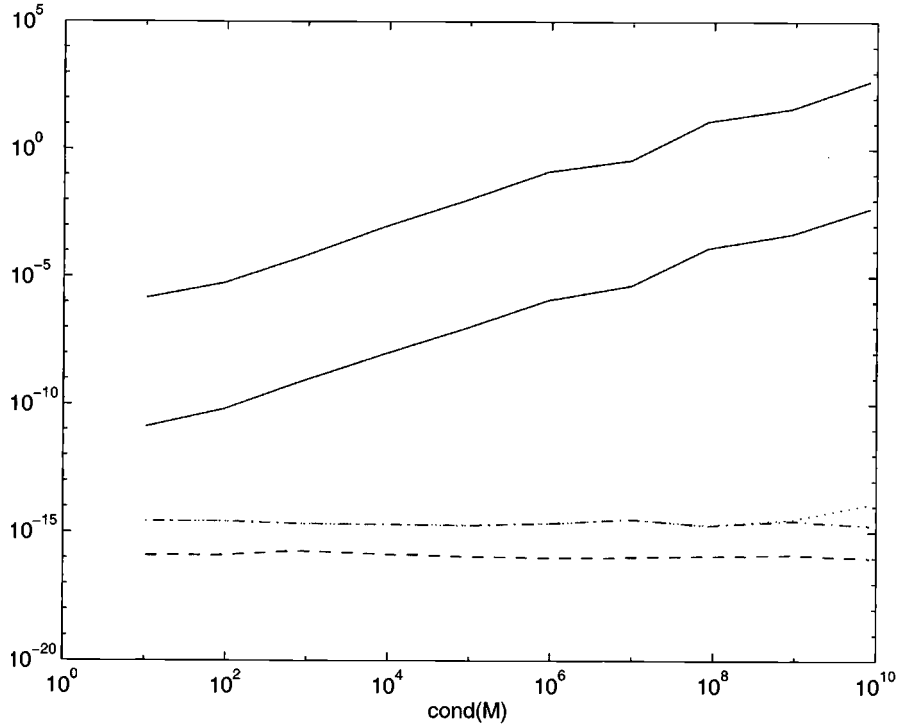


FIG. 5. *Error growth vs.* κ_M *for Method* 2.

be large and this will cause significant error. One might also expect that because the forward error in computing Z is necessarily of order $O(\kappa_{\mathbf{A}}\varepsilon)$, it would follow that no null-space method could provide accurate residuals.

The way forward, which is exploited in the analysis for Method 2, is that Method 2 determines a matrix $\widetilde{Z}$ for which $\widetilde{Z}^T A = O(\varepsilon)$. Thus, although the null space is inevitably badly determined when A is ill conditioned, Method 2 fixes on one particular basis matrix $\widetilde{Z}$ that is well behaved. This basis is an exact basis for an $O(\varepsilon)$ perturbation to A. Method 2 is able to solve this perturbed problem accurately. On the other hand, Method 1 essentially obtains a different approximation to Z for every solve with $\mathbf{A}$. Thus the computed reduced Hessian matrix $Z^T G Z$ does not correspond accurately to any one particular Z matrix.

In passing, it is interesting to remark that computing the factors

$$A = Q \begin{bmatrix} R \\ 0 \end{bmatrix} = [\, Q_1 \quad Q_2 \,] \begin{bmatrix} R \\ 0 \end{bmatrix},$$

and defining $Z = Q_2$, also provides a stable approach, not so much because it avoids the growth in Z (we have seen that this is rarely a problem), but because it also provides a fixed null-space reference basis, which is an exact basis for an $O(\varepsilon)$ perturbation to A.

In the context of quadratic programming, a common solution method for large sparse systems is to use some sort of product form method (Gauss–Jordan, Bartels–Golub–Reid, Forrest–Tomlin, etc. (see, for example, Suhl [12])). It is not clear that such methods provide $O(\varepsilon)$ solutions to the systems involving $\mathbf{A}$ that are solved

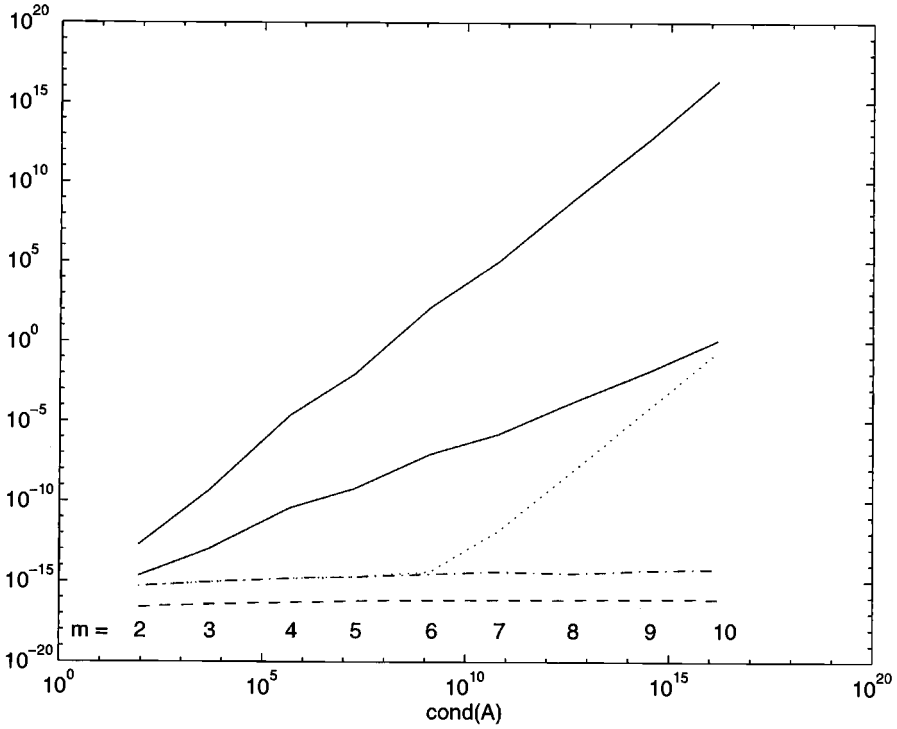


FIG. 6. *Error growth for Method* 2 *for a more ill-conditioned matrix.*

in Method 1 (although B–G–R may be stable in this respect). However, the main difficulty comes when the product form becomes too unwieldy and is reinverted. If A is ill conditioned, the refactorization of $\mathbf{A}$ is likely to determine a basis matrix Z that differs by $O(\kappa_{\mathbf{A}}\varepsilon)$ from that defined by the old product form. Thus the old reduced Hessian matrix Z^TGZ would not correspond accurately to that defined by the new Z matrix after reinversion. The only recourse would be to re-evaluate Z^TGZ on reinversion, which might be very expensive. Thus we do not see a product form method on its own as being suitable. Our paper has shown that if a fixed reference basis is generated, then accurate residuals are possible. We hope to show how this might be done in a subsequent paper by combining a product form method with another method such as LU factorization.

REFERENCES

[1] K. J. BATHE, *Finite Element Procedures in Engineering Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1982.

[2] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, Berlin, 1991.

[3] J. R. BUNCH AND L. C. 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] R. FLETCHER, *Factorizing symmetric indefinite matrices*, Linear Algebra Appl., 14 (1976), pp. 257–272.

[6] R. FLETCHER, *Practical Methods of Optimization*, 2nd ed., Wiley, Chichester, UK, 1987.

[7] A. Forsgren and W. Murray, *Newton methods for large-scale linear equality-constrained minimization*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 560–587.
[8] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright, *A Schur complement method for quadratic programming*, in Reliable Numerical Computation, M. G. Cox and S. Hammarling, eds., Clarendon Press, Oxford, 1990, pp. 113–138.
[9] N. I. M. Gould, *Constructing appropriate models for large-scale, linearly constrained, nonconvex, nonlinear optimization algorithms*, Tech. report RAL-TR-95-037, Rutherford Appleton Laboratory, Chilton, Oxfordshire, UK, 1995.
[10] N. J. Higham, *Stability of the Diagonal Pivoting Method with Partial Pivoting*, MCCM Numerical Analysis report 265, Manchester University, UK, 1995.
[11] G. W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.
[12] U. H. Suhl, *MOPS – Mathematical Optimization System*, European J. Oper. Res., 72 (1994), pp. 312–322.
[13] J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, NPL Notes in Applied Science 32, HMSO, London, 1963.
[14] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 4, pp. 959–980, October 1997

011

SPECTRAL PERTURBATION BOUNDS FOR POSITIVE DEFINITE MATRICES*

ROY MATHIAS†

Abstract. Let H and $H + \Delta H$ be positive definite matrices. It was shown by Barlow and Demmel and Demmel and Veselić that if one takes a componentwise approach one can prove much stronger bounds on $\lambda_i(H)/\lambda_i(H + \Delta H)$ and the components of the eigenvectors of H and $H + \Delta H$ than by using the standard normwise perturbation theory. Here a unified approach is presented that improves on the results of Barlow, Demmel, and Veselić. It is also shown that the growth factor associated with the error bound on the components of the eigenvectors computed by Jacobi's method grows linearly (rather than exponentially) with the number of Jacobi iterations required for convergence.

Key words. symmetric eigenvalue problem, positive definite matrix, graded matrix, perturbation theory, error analysis, Jacobi's method

AMS subject classifications. 65F15, 65G05, 15A18, 15A48

PII. S0895479894271081

1. Introduction.

If the positive definite matrix H can be written as $H = DAD$, where D is diagonal and A is much better conditioned than H, then the eigenvalues and eigenvectors of H are determined to a high relative accuracy if the entries of the matrix H are determined to a high relative accuracy. This was shown by Demmel and Veselić [2], building on work of Barlow and Demmel [1]. In this paper we strengthen some of the perturbation bounds in [2] and present a unified approach to proving these results. We also show that, just as conjectured in [2], the growth factor that arises in the bound on the accuracy of the components of the eigenvectors computed by Jacobi's method is linear rather than exponential.

We now give an outline of the paper and the main ideas in it and then define the notation. In section 2 we quickly reprove some of the eigenvalue and eigenvector perturbation bounds from [2] in a perhaps more unified way and derive bounds on the sensitivity of the eigenvalues to perturbations in any given entry of the matrix. The main idea in this section is that the analysis is reduced to standard perturbation theory if one can express additive perturbations as multiplicative perturbations. In this respect our approach is similar to that of Eisenstat and Ipsen in [4], except that they assume a multiplicative perturbation and then go on to derive bounds, whereas we assume an additive perturbation, which we rewrite as a multiplicative perturbation, before performing the analysis. Our results are the same as those in [4] for eigenvalues but not for eigenvectors. We briefly compare our approach to relative perturbation bounds with those in [1, 2, 4] in section 2.1. We also show that the relative gap associated with an eigenvalue is a very good measure of the distance (in the scaled norm) to the nearest matrix with a repeated eigenvalue.

In section 3 we consider the components of the eigenvectors of a graded positive

* Received by the editors July 13, 1994; accepted for publication (in revised form) by N. J. Higham October 23, 1996. This research was supported in part by National Science Foundation grant DMS-9201586 and much of it was done while the author was visiting the Institute for Mathematics and Its Applications at the University of Minnesota.

http://www.siam.org/journals/simax/18-4/27108.html

† Department of Mathematics, College of William and Mary, Williamsburg, VA 23187 (mathias@math.wm.edu).

definite matrix.[1] The key idea here is that if H is a graded positive definite matrix and U is orthogonal such that $H_1 = U^T H U$ is also graded, then U has a "graded" structure related to that of H and H_1.[2] This fact can be systematically applied to obtain componentwise perturbation bounds for the eigenvectors of graded positive definite matrices and componentwise bounds on the accuracy of the eigenvectors computed by Jacobi's method. The fact that the matrix of eigenvectors is "graded" has been observed in [1] and [2]; however, the results there were weaker than ours, and these papers did not exploit this graded structure to any great extent. The basic results on gradedness of eigenvectors are in section 3.1 and the applications are in section 3.2.

Let $M_{m,n}$ denote the space of $m \times n$ real matrices , and let $M_n \equiv M_{n,n}$. For a symmetric matrix H we let $\lambda_1(H) \geq \lambda_2(H) \geq \cdots \geq \lambda_n(H)$ denote its eigenvalues, ordered in decreasing order. For $X \in M_{m,n}$ we let $\sigma_1(X) \geq \sigma_2(X) \geq \cdots \geq \sigma_{\min\{m,n\}}(X)$ denote its singular values. The only norm that we use is the spectral norm (or 2-norm), and we denote it by $\|\cdot\|$, i.e., $\|X\| = \sigma_1(X)$. When we say that a matrix has unit columns we mean that its columns have unit Euclidean norm.

For a matrix or vector X, $|X|$ denotes its entrywise absolute value. For two matrices or vectors X and Y of the same dimensions we use $\min\{X, Y\}$ to denote their entrywise minimum, and we use $X \leq Y$ to mean that each entry of X is smaller than the corresponding entry of Y. To differentiate between the componentwise and positive semidefinite orderings we use $A \preceq B$ to mean that A and B are symmetric and $B - A$ is positive semidefinite. We use E to denote a matrix of ones and e to denote a column vector of ones—the dimension will be apparent from the context.

In studying the perturbation theory of eigenvectors we use the two notions of the *relative gap* between the eigenvalues that were introduced in [1], but we use different notation. Given a positive vector λ we define

$$\operatorname{relgap}(\lambda, i) = \min_{j \neq i} \frac{|\lambda_i - \lambda_j|}{\sqrt{\lambda_i \lambda_j}}$$

and

$$\operatorname{relgap}^*(\lambda, i) = \min_{j \neq i} \frac{|\lambda_i - \lambda_j|}{\lambda_i + \lambda_j}.$$

One similarity between the two relative gaps is that it is sufficient to take the minimum over $j = i - 1, i + 1$ in either case. However, it is easy to see that $\operatorname{relgap}^*(\lambda, i)$ is at most 1, while $\operatorname{relgap}(\lambda, i)$ can be arbitrarily large and that

$$\operatorname{relgap}(\lambda, i) \geq 2 \cdot \operatorname{relgap}^*(\lambda, i).$$

If $\lambda'_k = \lambda_k(1 + \alpha_k)$ where $|\alpha_k| \leq \delta$, then, as we show at the end of the section,

$$\operatorname{relgap}^*(\lambda', i) \geq \operatorname{relgap}^*(\lambda, i) - \delta. \tag{1.1}$$

Unfortunately the result for the perturbation to relgap is more complicated, and this sometimes complicates analysis and results involving relgap. (See [2, proof of Proposition 2.6] for such an instance.)

[1] We say that the positive definite matrix H is graded if $H = DAD$, where D is diagonal and A is much better conditioned than H.

[2] By this we mean that both $\|D^{-1}UD_1\|$ and $\|DUD_1^{-1}\|$ are not much larger than 1, where D and D_1 are diagonal matrices such that the diagonal elements of $D^{-1}HD^{-1}$ and $D_1^{-1}H_1D_1^{-1}$ are all 1. We use quotes because this is not the usual definition of gradedness, but, nonetheless it is related to the gradedness of H and H_1.

It is not clear which relative gap one should use, or whether one should use both, or perhaps the relative gap used in [4]. In [2] it was suggested that $\mathrm{relgap}(\lambda(H), i)$ is the appropriate measure of the relative gap between $\lambda_i(H)$ and the rest of the eigenvalues of H and that $\mathrm{relgap}(\sigma(G), i)$ is the appropriate measure of the relative gap between $\sigma_i(G)$ and the rest of the singular values of G. The eigenvector results in Theorems 3.5 and 2.9 and Corollary 2.10 and the singular vector results in Theorem 2.8 suggest that this is not the case.

Luckily, one is most interested in the relative gap when it is small, and in this case it doesn't make much difference which definition one chooses. For example, if $\mathrm{relgap}(\lambda, i) \leq 1.5$, then one can check that

$$2 \cdot \mathrm{relgap}^*(\lambda, i) \leq \mathrm{relgap}(\lambda, i) \leq 2.5\ \mathrm{relgap}^*(\lambda, i). \tag{1.2}$$

One can also check that the left-hand inequality is always valid by a simple application of the arithmetic–geometric mean inequality.

Let us now prove (1.1). Define f on $(0, \infty)^2$ by

$$f(x_1, x_2) = \frac{|x_1 - x_2|}{x_1 + x_2}.$$

Then

$$\mathrm{relgap}^*(\lambda, i) = \max_{j \neq i} f(\lambda_i, \lambda_j).$$

So in order to prove (1.1) it is sufficient to prove that for any $\lambda_1, \lambda_2, \alpha_1, \alpha_2, \delta$ for which

$$\max\{|\alpha_1|, |\alpha_2|\} \leq \delta \leq \frac{|\lambda_1 - \lambda_2|}{\lambda_1 \lambda_2}, \tag{1.3}$$

we must have

$$f(\tilde{\lambda}_1, \tilde{\lambda}_2) \geq f(\lambda_1, \lambda_2) - \delta. \tag{1.4}$$

Without loss of generality $\lambda_1 > \lambda_2$. The bound (1.3) implies that $\tilde{\lambda}_1 \geq \tilde{\lambda}_2$. Since $\lambda_1 \geq \lambda_2$ and $\tilde{\lambda}_1 \geq \tilde{\lambda}_2$ it follows that

$$f(\lambda_1, \lambda_2) = \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2} = 1 - \frac{2\lambda_2}{\lambda_1 + \lambda_2}, \tag{1.5}$$

$$f(\tilde{\lambda}_1, \tilde{\lambda}_2) = \frac{\tilde{\lambda}_1 - \tilde{\lambda}_2}{\tilde{\lambda}_1 + \tilde{\lambda}_2} = 1 - \frac{2\tilde{\lambda}_2}{\tilde{\lambda}_1 + \tilde{\lambda}_2}. \tag{1.6}$$

In writing (1.6) as

$$f(\tilde{\lambda}_1, \tilde{\lambda}_2) = 1 - 2\left(\frac{\tilde{\lambda}_1}{\tilde{\lambda}_2} + 1\right)^{-1}$$

one sees that $f(\tilde{\lambda}_1, \tilde{\lambda}_2)$, thought of as a function of α_1 and α_2, is minimized when $\alpha_1 = -\delta$ and $\alpha_2 = \delta$. Substituting these values for α_1 and α_2 and substituting the expressions (1.5) and (1.6) in (1.4), we see that it is sufficient to prove

$$-\frac{2\lambda_2}{\lambda - 1 + \lambda_2} + \frac{2\lambda_2(1 + \delta)}{\lambda_1(1 - \delta) + \lambda_2(1 + \delta)} \leq \delta \tag{1.7}$$

or, equivalently,

$$\frac{4\lambda_1\lambda_2\delta}{(\lambda-1+\lambda_2)(\lambda_1(1-\delta)+\lambda_2(1+\delta))} \le \delta, \tag{1.8}$$

which is equivalent to

$$\frac{4\lambda_1\lambda_2}{(\lambda-1+\lambda_2)(\lambda_1(1-\delta)+\lambda_2(1+\delta))} \le 1. \tag{1.9}$$

The left-hand side of (1.9) is an increasing function of δ, and so in order to verify (1.9) it is sufficient to verify it when δ is as large as possible—that is, when

$$\delta = \frac{\lambda_1-\lambda_2}{\lambda_1+\lambda_2}.$$

Straightforward algebra shows that (1.9) holds with equality when one substitutes this value of δ. Thus we have verified (1.1). The bound (1.1) is a slight improvement over [7, Proposition 3.3, equation (3.8)] in the case $p = 1$.

2. A unified approach. In this section we give a unified approach to some of the inequalities in [2] and [1]. This approach also allows one to bound the relative perturbation in the eigenvalues of a positive definite matrix caused by a perturbation in a particular entry.

The key idea in this section is to express the additive perturbation $H + \Delta H$ as a multiplicative perturbation of H. Given a multiplicative perturbation of a matrix it is quite natural that the perturbation of the eigenvalues and eigenvectors is also multiplicative. It is then a small step from this multiplicative perturbation to the componentwise perturbation bounds that we desire. There are two ways to write $H + \Delta H$ as a multiplicative perturbation:

$$H + \Delta H = [(H+\Delta H)^{\frac{1}{2}} H^{-\frac{1}{2}}]\, H\, [(H+\Delta H)^{\frac{1}{2}} H^{-\frac{1}{2}}]^T \tag{2.10}$$

and

$$H + \Delta H = Y(I + Y^{-1}(\Delta H)Y^{-T})Y^T, \tag{2.11}$$

where $H = YY^T$. (One possible choice of Y is $H^{\frac{1}{2}}$.) If one wants to prove eigenvalue inequalities it seems that both representations give the same bounds. If one uses the representation (2.10), then Ostrowski's theorem [6, Theorem 4.5.9] yields the relation between the eigenvalues of H and $H + \Delta H$—this is the route taken in [4]. We shall use (2.11) and the monotonicity principle (Theorem 2.1) because the proofs are slightly quicker. Demmel and Veselić [2] and Barlow and Demmel [1] used the Courant–Fisher min-max representation of the eigenvalues of a Hermitian matrix to derive similar results.

In Jacobi's method one encounters positive definite matrices $H = DAD$ and $\Delta H = D(\Delta A)D$, where D is diagonal and A can be much better conditioned than H. For this reason Demmel and Veselić assumed the matrices $H = DAD$ and $\Delta H = D(\Delta A)D$ with D diagonal to be the data in their work [2]. We consider a slightly more general situation and just assume that H and $H + \Delta H$ are positive definite. We consider this more general setting first to show that one can prove relative perturbation bounds for positive definite matrices without assuming that the matrices are graded and second because the results are slightly cleaner in the general case. (For example,

the statement of Theorem 2.9, which deals with the general case, is cleaner than the statement of Corollary 2.10, which deals with the special case where D is diagonal.) Lemma 2.2 allows us to derive their results as corollaries of ours.

THEOREM 2.1 (Monotonicity Principle [6, Corollary 4.3.3]). *Let $A, B \in M_n$. If $A \preceq B$, then*

$$\lambda_i(A) \le \lambda_i(B), \quad i = 1, \ldots, n.$$

The following lemma will be useful in applying our general results in special situations.

LEMMA 2.2. *Let H be positive definite and let ΔH be arbitrary. Let $Y \in M_n$ be such that $H = YY^T$. Then*

$$\|Y^{-1}(\Delta H)Y^{-T}\| = \|H^{-\frac{1}{2}}(\Delta H)H^{-\frac{1}{2}}\|.$$

Furthermore, if $H = DAD^T$ and $\Delta H = D(\Delta A)D^T$, then

$$\|H^{-\frac{1}{2}}(\Delta H)H^{-\frac{1}{2}}\| = \|A^{-\frac{1}{2}}(\Delta A)A^{-\frac{1}{2}}\|.$$

Proof. Since $YY^T = H^{\frac{1}{2}}(H^{\frac{1}{2}})^T$ there must be an orthogonal matrix Q such that $Y = H^{\frac{1}{2}}Q$. Thus

$$\|Y^{-1}\Delta H Y^{-T}\| = \|Q^T H^{-\frac{1}{2}}\Delta H H^{-\frac{1}{2}}Q\| = \|H^{-\frac{1}{2}}\Delta H H^{-\frac{1}{2}}\|.$$

For the second part of the lemma take $Y = DA^{\frac{1}{2}}$ and apply the first part. Then we have

$$\|H^{-\frac{1}{2}}(\Delta H)H^{-\frac{1}{2}}\| = \|A^{-\frac{1}{2}}D^{-1}D(\Delta A)D^T D^{-T}A^{-\frac{1}{2}}\| = \|A^{-\frac{1}{2}}(\Delta A)A^{-\frac{1}{2}}\|,$$

as required. □

Note that if D is diagonal, as it will be in applications, then $D = D^T$. Also, using the notation of Lemma 2.2 we have

$$\eta = \|A^{-\frac{1}{2}}(\Delta A)A^{-\frac{1}{2}}\| \le \|A^{-\frac{1}{2}}\|\,\|\Delta A\|\,\|A^{-\frac{1}{2}}\| = \|A^{-1}\|\,\|\Delta A\|.$$

Our bounds are in terms of η while those of Demmel and Veselić in [2] are in terms of the larger quantity $\|A^{-1}\|\,\|\Delta A\|$. They assumed that the diagonal elements of A are all 1. This is not always necessary, though it is a good choice of A in that it approximately minimizes $\|A^{-1}\|\,\|\Delta A\|$. We only assume that the diagonal elements of A are 1 when it is necessary.

2.1. Eigenvalues and singular values. Here is our main eigenvalue perturbation theorem.

THEOREM 2.3. *Let H, $H+\Delta H \in M_n$ be positive definite and let $\eta = \|H^{-\frac{1}{2}}\Delta H H^{-\frac{1}{2}}\|$. Then*

$$(1-\eta)\lambda_i(H) \le \lambda_i(H+\Delta H) \le (1+\eta)\lambda_i(H).$$

Proof. Write $H + \Delta H = H^{\frac{1}{2}}(I + H^{-\frac{1}{2}}(\Delta H)H^{-\frac{1}{2}})H^{\frac{1}{2}}$. Since

$$-\eta I \le H^{-\frac{1}{2}}(\Delta H)H^{-\frac{1}{2}} \le \eta I$$

we have

$$(1-\eta)H \le H + \Delta H \le (1+\eta)H.$$

The monotonicity principle (Theorem 2.1) now gives the required bounds. □

Using the second part of Lemma 2.2 we obtain a result that is essentially the same as [2, Theorem 2.3].

THEOREM 2.4. *Let $H = DAD$, $\hat{H} = D(A+\Delta A)D \in M_n$ be positive definite, assume that D diagonal, and let $\eta = \|A^{-\frac{1}{2}}\Delta A A^{-\frac{1}{2}}\|$. Then*

$$1-\eta \;\le\; \frac{\lambda_i(\hat{H})}{\lambda_i(H)} \;\le\; 1+\eta, \qquad i=1,\dots,n. \tag{2.12}$$

As another corollary of the monotonicity principle we have a useful relation between the diagonal elements of a positive definite matrix and its eigenvalues [2, Proposition 2.10].

COROLLARY 2.5. *Let $H = DAD \in M_n$ be a positive definite matrix and assume that D is diagonal and that the main diagonal entries of A are all 1 while the main diagonal entries of H are ordered in decreasing order. Then*

$$\lambda_n(A) \le \frac{\lambda_i(H)}{h_{ii}} \le \lambda_1(A), \;\; i=1,\dots,n. \tag{2.13}$$

Proof. Since $\lambda_n(A)I \le A \le \lambda_1(A)I$ it follows that $\lambda_n(A)D^2 \le DAD \le \lambda_1(A)D^2$. The matrix D^2 is diagonal so its eigenvalues are its diagonal elements and these are h_{ii}, $1,\dots,n$. The result now follows from the monotonicity principle. □

One would expect that the eigenvalues of H are more sensitive to perturbations in some entries of H and less sensitive to perturbations in others. Stating the bound in terms of $\eta = \|A^{-\frac{1}{2}}(\Delta A)A^{-\frac{1}{2}}\|$ allows one to derive stronger bounds on the sensitivity of the eigenvalues of H to a perturbation in any one of the entries (or two corresponding off-diagonal entries) of H than if we had replaced η by $\|\Delta A\|\,\|A^{-1}\|$. Let us assume the notation of the theorem. Let $E_{ij} = e_i e_j^T$ (e_i is the unit n-vector with ith component equal to 1). Suppose that $\Delta A = \epsilon E_{jj}$, that is a relative perturbation of ϵ in the jth main diagonal entry, then

$$\|A^{-\frac{1}{2}}(\Delta A)A^{-\frac{1}{2}}\| = \|\epsilon A^{-\frac{1}{2}} e_j e_j^T A^{-\frac{1}{2}}\| = |\epsilon|\,\|e_j^T A^{-\frac{1}{2}} A^{-\frac{1}{2}} e_j\| = |\epsilon|(A^{-1})_{jj},$$

and so

$$1-|\epsilon|(A^{-1})_{jj} \;\le\; \frac{\lambda_i(H+\Delta H)}{\lambda_i(H)} \;\le\; 1+|\epsilon|(A^{-1})_{jj}, \;\; i=1,\dots,n. \tag{2.14}$$

In fact, we can say more. If $\epsilon > 0$, then $H \le H + \Delta H$, and so from the monotonicity principle we know that $\lambda_i(H) \le \lambda_i(H+\Delta H)$, and so the lower bound in (2.14) can be taken as 1, and vice versa if $\epsilon < 0$. If $\|A^{-1}\| \gg (A^{-1})_{jj}$, as is quite possible for some values of j, then the bound (2.14) is much better than (2.12) with η replaced by $\|A^{-1}\|\,\|\Delta A\|$.

If $\Delta A = \epsilon(E_{ij} + E_{ji})$, a symmetric perturbation in entries ij and ji, then for any $\alpha > 0$

$$-|\epsilon|(\alpha E_{ii} + \alpha^{-1}E_{jj}) \preceq \Delta A \preceq |\epsilon|(\alpha E_{ii} + \alpha^{-1}E_{jj}).$$

Now taking $\alpha = \sqrt{(A^{-1})_{jj}/(A^{-1})_{ii}}$ we have

$$\begin{aligned}\|A^{-\frac{1}{2}}\Delta A A^{-\frac{1}{2}}\| &\le |\epsilon|\ \|A^{-\frac{1}{2}}(\alpha E_{ii} + \alpha^{-1}E_{jj})A^{-\frac{1}{2}}\| \\ &\le |\epsilon|\ (\alpha(A^{-1})_{ii} + \alpha^{-1}(A^{-1})_{jj}) \\ &= 2\sqrt{(A^{-1})_{ii}(A^{-1})_{jj}}\ |\epsilon|\end{aligned}$$

and so for $i = 1, \ldots, n$

$$1 - 2|\epsilon|\sqrt{(A^{-1})_{ii}(A^{-1})_{jj}} \ \le\ \frac{\lambda_i(\hat{H})}{\lambda_i(H)}\ \le\ 1 + 2|\epsilon|\sqrt{(A^{-1})_{ii}(A^{-1})_{jj}}. \tag{2.15}$$

One may hope to prove a bound with $|(A^{-1})_{ij}|$ instead of $[(A_{ii})^{-1}(A_{jj})^{-1}]^{\frac{1}{2}}$. To see that such a bound is not possible consider the case $A = I$. Then the off-diagonal elements of A^{-1} are 0, but clearly perturbing an off-diagonal element of A does change the eigenvalues of DAD.

One can obtain similar bounds on the perturbation of the eigenvectors, singular values, and singular vectors caused by a perturbation in one of the elements of the matrix. In the case of eigenvectors and singular vectors one can obtain normwise and componentwise bounds. The bounds for singular values and singular vectors involve a row of B^{-1} (or $B^{\dagger}$ if $B \in M_{m,n}$ and B is of full rank) rather than just one element of the inverse (or pseudoinverse).

2.2. Eigenvectors and singular vectors. Now let us see how this approach gives normwise perturbation bounds for the eigenvectors of a graded positive definite matrix in terms of the relative gap between the eigenvalues. Let H be positive definite. Let U be an orthogonal matrix with the jth column an eigenvector of H corresponding to $\lambda_j(H)$, and let Λ be a diagonal matrix with ii element $\lambda_i(H)$. Then

$$H + \Delta H = U\Lambda^{\frac{1}{2}}(I + \Delta)\Lambda^{\frac{1}{2}}U^T,$$

where $\Delta = Y^{-1}(\Delta H)Y^{-T}$ and $Y = U\Lambda^{\frac{1}{2}}$. Since $YY^T = H$, the first part of Lemma 2.2 implies that

$$\|\Delta\| = \|H^{-\frac{1}{2}}(\Delta H)H^{-\frac{1}{2}}\| \equiv \eta. \tag{2.16}$$

Let $\hat{u}$ be an eigenvector of $\Lambda^{\frac{1}{2}}(I + \Delta)\Lambda^{\frac{1}{2}}$. Then $\tilde{u} = U\hat{u}$ is an eigenvector of $H + \Delta H$. The vector $u = Ue_j$ is an eigenvector of H, and so the normwise difference between u and $\tilde{u}$ is

$$\|u - \tilde{u}\| = \|U(e_j - \hat{u})\| = \|e_j - \hat{u}\|. \tag{2.17}$$

So to show that $\tilde{u}$ can be chosen such that $\|u - \tilde{u}\|$ is small we must show that $\hat{u}$ can be chosen to be close to e_j. We do this in Lemma 2.6, which follows easily from the standard perturbation theory given in [5, pp. 345–346].

We have used the fact that U is orthogonal in (2.17), and hence has norm 1, to obtain a normwise bound on $u - \tilde{u}$. In section 3.2 we use the componentwise bounds (2.18–2.19) on U to derive a componentwise bound on $u - \tilde{u}$.

LEMMA 2.6. *Let $\Lambda = \mathrm{diag}(\lambda)$ have main diagonal elements ordered in decreasing order and assume that $\lambda_{j+1} < \lambda_j < \lambda_{j-1}$. Let X be a symmetric matrix and let*

$H(\epsilon) = \Lambda + \epsilon X$. Then for ϵ sufficiently small $\lambda_j(\epsilon) = \lambda_j(H(\epsilon))$ is distinct, and one can choose $\hat{u}(\epsilon)$ to be an eigenvector of $H(\epsilon)$ such that

$$\hat{u}(\epsilon)_j = 1 + O(\epsilon^2), \tag{2.18}$$

$$|\hat{u}(\epsilon)_i| \le \epsilon \frac{|x_{ij}|}{|\lambda_i - \lambda_j|} + O(\epsilon^2), \quad i \ne j, \tag{2.19}$$

and so

$$\|\hat{u}(\epsilon) - e_j\| \le \epsilon \left(\sum_{i \ne j} \frac{|x_{ij}|^2}{|\lambda_i - \lambda_j|^2} \right)^{\frac{1}{2}} + O(\epsilon^2). \tag{2.20}$$

If we take $X = \Lambda^{\frac{1}{2}} \Delta \Lambda^{\frac{1}{2}}$ in Lemma 2.6, then one can see that the coefficient of ϵ on the right-hand side of (2.20) is bounded by

$$\left(\sum_{i \ne j} \frac{|\delta_{ij}|^2 \lambda_i \lambda_j}{|\lambda_i - \lambda_j|^2} \right)^{\frac{1}{2}} \le \operatorname{relgap}^{-1}(\lambda, j) \left(\sum_{i \ne j} |\delta_{ij}|^2 \right)^{\frac{1}{2}} \le \operatorname{relgap}^{-1}(\lambda, j) \|\Delta\|,$$

where δ_{ij} is the ij element of Δ. Substituting η for $\|\Delta\|$ from (2.16) we get

$$\|\hat{u} - e_j\| \le \frac{\eta\, \epsilon}{\operatorname{relgap}(\lambda, j)} + O(\epsilon^2).$$

From (2.17) it follows that we have the same bound on $\|u - \tilde{u}\|$.

We summarize the argument in the following theorem.

THEOREM 2.7. *Let $H \in M_n$ be positive definite and let $\|H^{-\frac{1}{2}} \Delta H H^{-\frac{1}{2}}\|$. Let $H(\epsilon) = H + \epsilon \Delta H$. Let $\lambda_j(\epsilon) = \lambda_j(H(\epsilon))$. Assume that $\lambda_j(0)$ is a simple eigenvalue of H. Let u be a corresponding unit eigenvector of H. Then, for sufficiently small ϵ, there is an eigenvector $u(\epsilon)$ of $H(\epsilon)$ corresponding to $\lambda_j(\epsilon)$ such that*

$$\|u - u(\epsilon)\| \le \frac{\epsilon \|H^{-\frac{1}{2}} (\Delta H) H^{-\frac{1}{2}}\|}{\operatorname{relgap}(\lambda, j)} + O(\epsilon^2). \tag{2.21}$$

As mentioned earlier, we may replace η by $\|A^{-1}\| \|\Delta A\|$. The resulting bound improves the bound in [2, Theorem 2.5] by a factor of $\sqrt{n-1}$.

Eisenstat and Ipsen also give a bound on the perturbation of eigenvectors which involves a relative gap [4, Theorem 2.2]. Their bound relates the eigenvectors of H and those of KHK^T, where $K \in M_n$ is nonsingular. It is an absolute bound—not a first-order bound. To obtain a bound of the form (2.21) from [4, Theorem 2.2] one must find a bound on $\|(H + \epsilon \Delta H)^{\frac{1}{2}} H^{-\frac{1}{2}} - I\|$ of the form

$$\|(H + \epsilon \Delta H)^{\frac{1}{2}} H^{-\frac{1}{2}} - I\| \le c\epsilon \|H^{-\frac{1}{2}} (\Delta H) H^{-\frac{1}{2}}\| + O(\epsilon^2). \tag{2.22}$$

It is shown in [9] that if (2.22) is to hold for all $n \times n$ H and ΔH with H positive definite, then the constant c must depend on n and must grow like $\log n$. That is, a direct application of [4, Theorem 2.2] to the present situation does not yield (2.21). However, one can derive (2.21) using the idea behind the proof of [4, Theorem 2.2] and a more careful argument [3].

Veselić and the author have used ideas similar to those in this section to prove a nonasymptotic relative perturbation bound on the eigenvectors of a positive definite matrix [13].

One can apply Lemma 2.6 to GG^T and G^TG and thereby remove the factor $\sqrt{n-.5}$ from the bound on the perturbation of the right and left singular vectors given in [2, Theorem 2.16]. Note that one must apply Lemma 2.6 directly in order to obtain the strongest result. If one applies Theorem 2.7 to G^TG the resulting bound contains an extra factor $(1-\|\Delta GG^\dagger\|)^{-1}$. Notice that the bound on the right and left singular vectors is not the same—the bound on the right singular vectors is potentially much smaller since relgap can be much larger than relgap*.

THEOREM 2.8. *Let $G, G+\Delta G \in M_{m,n}$ and let $G^\dagger$ be the pseudoinverse of G. Assume that G is of rank $q = \min\{m,n\}$ and that $\Delta G = \Delta GG^\dagger G$.*

Let $G(\epsilon) = G + \epsilon\Delta G$. Take j between 1 and q and assume that $\sigma_j(G)$ is simple. Let u and v be left and right singular vectors of G corresponding to $\sigma_j(G)$. Then for sufficiently small ϵ, there are left and right singular vectors of $G(\epsilon)$, $u(\epsilon)$, and $v(\epsilon)$ corresponding to $\sigma_j(G(\epsilon))$ such that

$$\|u - u(\epsilon)\| \le \frac{2\|\Delta GG^\dagger\|\epsilon}{\text{relgap}(\sigma^2(G), j)} + O(\epsilon^2), \tag{2.23}$$

$$\|v - v(\epsilon)\| \le \frac{\sqrt{2}\|\Delta GG^\dagger\|\epsilon}{\text{relgap}^*(\sigma^2(G), j)} + O(\epsilon^2). \tag{2.24}$$

Proof. Let $U\Sigma V^T$ be a singular value decomposition of G; here u and V are square and Σ is rectangular. First let us consider the right singular vectors, which are the eigenvectors of G^TG.

$$\begin{aligned} G^T(\epsilon)G(\epsilon) &= G^T(I+\epsilon\Delta GG^\dagger)^T(I+\epsilon\Delta GG^\dagger)G \\ &= V\Sigma^T(I+\epsilon(U^TG^{\dagger T}\Delta G^TU + U^T\Delta GG^\dagger U))\Sigma V^T + O(\epsilon^2) \\ &= V\Sigma^T(I+\epsilon F)\Sigma V^T + O(\epsilon^2), \end{aligned}$$

where $F = U^TG^{\dagger T}\Delta G^TU + U^T\Delta GG^\dagger U$ and hence has norm at most $2\|\Delta GG^\dagger\| = 2\eta$. Now from (2.20) one can choose $\tilde{u}$ as a jth eigenvector of $\Sigma^T(I+\epsilon F)\Sigma$ that differs in norm from e_j by at most

$$\epsilon\left(\sum_{i\neq j}\frac{|f_{ij}|^2\sigma_i^2\sigma_j^2}{|\sigma_i^2-\sigma_j^2|^2}\right)^{\frac{1}{2}} \le \text{relgap}^{-1}(\sigma^2, j)\|F\|\epsilon \le 2\text{relgap}^{-\frac{1}{2}}(\sigma^2, j)\eta\epsilon$$

to first order in ϵ. Hence, we have the same bound on $\|u - V\tilde{u}\|$ to first order in ϵ. The vector $V\tilde{u}$ is an eigenvector (corresponding to jth eigenvalue) of

$$V\Sigma^T(I+\epsilon F)\Sigma V^T,$$

which is equal to $G^T(\epsilon)G(\epsilon)$ up to $O(\epsilon^2)$ terms. Since the jth singular value of $G(\epsilon)$ is simple, it follows that $V\tilde{u}$ is a right singular vector of of $G(\epsilon)$ up to $O(\epsilon^2)$.

Now let us consider the left singular vectors. As above we can show that

$$G(\epsilon)G(\epsilon)^T = U(\Sigma\Sigma^T + \epsilon(F\Sigma\Sigma^T + \Sigma\Sigma^TF^T))U^T + O(\epsilon^2),$$

where $F = (\Delta G)G^\dagger U$ and has norm at most η. So by (2.20) there is an eigenvector of $\Sigma\sigma^T + \epsilon(F\Sigma\Sigma^T + \Sigma\Sigma^T F)$ that differs from e_j in norm by at most

$$\begin{aligned}\epsilon\left(\sum_{i\neq j}\frac{(\sigma_i f_{ij}+\sigma_j f_{ji})^2}{(\sigma_i^2-\sigma_j^2)^2}\right)^{\frac{1}{2}} &\leq \epsilon\left(\sum_{i\neq j}\frac{(\sigma_i^2+\sigma_j^2)(f_{ij}^2+f_{ji}^2)}{(\sigma_i^2-\sigma_j^2)^2}\right)^{\frac{1}{2}}\\ &\leq \epsilon\left(\max_{i\neq j}\frac{\sigma_i^2+\sigma_j^2}{|\sigma_i^2-\sigma_j^2|^2}\right)^{\frac{1}{2}}\left(\sum_{i\neq j} f_{ij}^2+f_{ji}^2\right)^{\frac{1}{2}}\\ &\leq \mathrm{relgap}^{*-1}(\sigma^2,j)\eta\,\epsilon\end{aligned}$$

to first order in ϵ. In the same way as before, we can now deduce that there is a vector $v(\epsilon)$ with this distance of v. □

2.3. Distance to nearest ill-posed problem. It was shown in [1, Proposition 9] that $\mathrm{relgap}(\lambda(H), i)$ is approximately the distance from H to the nearest matrix with a multiple ith eigenvalue in the case that H is a scaled diagonally dominant symmetric matrix and distances are measured with respect to the grading of H. We show that there is a similar result for positive definite matrices. In Theorem 2.9 we show that $\mathrm{relgap}(\lambda(H), i)$ is *exactly* the distance to the nearest matrix with a repeated ith eigenvalue when we use the norm $N(X) = \|H^{-\frac{1}{2}}XH^{-\frac{1}{2}}\|$. We strengthen [1, Proposition 9] in Corollary 2.10 — our upper and lower bounds on the distance differ by a factor of $\kappa(A)$ while those in [1, Proposition 9] differ by a factor of about $\kappa^4(A)$, a potentially large difference. Our bound is considerably simpler than that in [1]; it doesn't involve factors of n (although one could replace $\lambda_1(A)$ by n) and its validity doesn't depend on the value of the relative gap (the bound in [1] has the requirement $\mathrm{relgap} \leq \frac{1}{2}$). Block diagonal examples show that not every eigenvalue of H will have the maximum sensitivity $\lambda_n^{-1}(A)$ and so this difference in the upper and lower bounds is to be expected. That is to say that one cannot hope to improve the bound (2.26) by more than a factor of $\lambda_1(A) \leq n$. Our bound involves relgap^* while the bound in [1] involves relgap. All these reasons suggest that relgap^{*-1}, and not relgap^{-1}, is the right measure of the distance to the nearest problem with a repeated ith eigenvalue.

THEOREM 2.9. *Let H be positive definite. Let $\lambda_i(H)$ be a simple eigenvalue of H, so that* $\mathrm{relgap}^*(\lambda_i(H)) > 0$. *Let*

$$\delta = \min\{\|H^{-\frac{1}{2}}(\Delta H)H^{-\frac{1}{2}}\| : \lambda_i(H+\Delta H) \text{ is a multiple eigenvalue of } H+\Delta H\}.$$

Then

$$\delta = \mathrm{relgap}^*(\lambda(H), i).$$

Proof. First we show that $\delta \geq \mathrm{relgap}^*(\lambda(H), i)$. Let ΔH be a perturbation that attains the minimum in the definition of δ. Then $\delta = \|H^{-\frac{1}{2}}(\Delta H)H^{-\frac{1}{2}}\|$. Let $\lambda_k = \lambda_k(H)$ and $\lambda'_k = \lambda_k(H+\Delta H)$ for $k = 1, \ldots, n$. By Theorem 2.3 we know that

$$\lambda'_k = \lambda_k(1+\alpha_k), \quad \text{where } |\alpha_k| \leq \delta. \tag{2.25}$$

Since ΔH has a multiple ith eigenvalue there is an index $j \neq i$ for which $\lambda'_i = \lambda'_j$. By (2.25) we must have $\lambda_j(1+\alpha_j) = \lambda_i(1+\alpha_i)$ and so

$$|\lambda_i - \lambda_j| \leq (\lambda_i + \lambda_j)\delta,$$

which implies

$$\operatorname{relgap}^*(\lambda(H), i) \leq \frac{|\lambda_i - \lambda_j|}{\lambda_i + \lambda_j} \leq \delta.$$

Now we show that $\delta = \operatorname{relgap}^*(\lambda(H), i)$ is attainable. Choose a value j such that $j \in \{i-1, i+1\}$

$$\frac{|\lambda_i - \lambda_j|}{\lambda_i + \lambda_j} = \operatorname{relgap}^*(\lambda, i).$$

(One can easily show that this is possible.) Set $\alpha = \lambda_j/(\lambda_i + \lambda_j)$ and take

$$\Delta H = (\lambda_i - \lambda_j)[\alpha x_j x_j^T - (1-\alpha) x_i x_i^T],$$

where x_i and x_j are unit eigenvectors of H corresponding to λ_i and λ_j. One can check that $\lambda_i(H + \Delta H) = \lambda_j(H + \Delta H)$. Because x_i and x_j are eigenvectors of $H^{-\frac{1}{2}}$

$$H^{-\frac{1}{2}}(\Delta H)H^{-\frac{1}{2}} = (\lambda_i - \lambda_j)[\alpha \lambda_j^{-1} x_j x_j^T - (1-\alpha)\lambda_i^{-1} x_i x_i^T].$$

Because x_i and x_j are orthogonal it follows that

$$\|H^{-\frac{1}{2}}(\Delta H)H^{-\frac{1}{2}}\| = |\lambda_i - \lambda_j| \cdot \max\{\alpha\lambda_j^{-1}, (1-\alpha)\lambda_i^{-1}\} = \frac{|\lambda_i - \lambda_j|}{\lambda_i + \lambda_j}$$

as required. □

COROLLARY 2.10. *Let $H = DAD \in M_n$ be positive definite and assume that D is diagonal and that the main diagonal entries of A are 1. Let $\lambda_i(H)$ be a simple eigenvalue of H, so that* $\operatorname{relgap}^*(\lambda_i(H)) > 0$. *Let*

$$\delta_D = \min\{\|D^{-1}(\Delta H)D^{-1}\| : \lambda_i(H + \Delta H) \text{ is a multiple eigenvalue of } H + \Delta H\}.$$

Then

$$\lambda_n(A) \operatorname{relgap}^*(\lambda(H), i) \leq \delta \leq \lambda_1(A) \cdot \operatorname{relgap}^*(\lambda(H), i). \tag{2.26}$$

Proof. Because

$$\lambda_n(A)\|A^{-\frac{1}{2}}D^{-1}\Delta H D^{-1}A^{-\frac{1}{2}}\| \leq \|D^{-1}\Delta H D^{-1}\| \leq \lambda_1(A)\|A^{-\frac{1}{2}}D^{-1}\Delta H D^{-1}A^{-\frac{1}{2}}\|$$

and because, by Lemma 2.2, we have

$$\|A^{-\frac{1}{2}}D^{-1}\Delta H D^{-1}A^{-\frac{1}{2}}\| = \|H^{-\frac{1}{2}}\Delta H H^{-\frac{1}{2}}\|$$

it follows that

$$\lambda_n(A)\delta \leq \delta_D \leq \lambda_1(A)\delta.$$

The result now follows from Theorem 2.9. □

3. Eigenvector components. It was shown in [1] that the eigenvectors of a scaled diagonally dominant matrix are scaled in the same way as the matrix. Essentially the same proof yields [2, Proposition 2.8]. We strengthen these by a factor $\kappa(A)$ in Corollaries 3.2 and 3.3. In section 3.2 we strengthen many of the results in [2] by using the stronger results in section 3.1 and show that the growth factor in the error bound on the eigenvectors computed by Jacobi's method is linear rather than exponential (Theorem 3.8). We also give improved componentwise bounds for the perturbation of singular vectors (Theorems 3.6 and 3.7). It is essential that the D_i be diagonal in this section as we are considering the components of the eigenvectors.

3.1. Gradedness of eigenvectors. Here we give some simple results on the graded structure of an orthogonal matrix that transforms one graded positive definite matrix into another and use this to derive results on the eigenvectors of a graded positive definite matrix.

LEMMA 3.1. *Let $H_1 = C^T H_0 C$ and let $H_i = D_i A_i D_i$, where the main diagonal entries of the A_i are 1 and the D_i are diagonal. Assume that $H_0 \in M_n$ and $H_1 \in M_m$ are positive definite. Then*

$$\|D_0 C D_1^{-1}\| \le \lambda_1^{\frac{1}{2}}(A_1)\lambda_n^{-\frac{1}{2}}(A_0) \le \sqrt{m}\ \lambda_n^{-\frac{1}{2}}(A_0). \tag{3.27}$$

Proof. It is easy to check that

$$A_1 = D_1^{-1} H_1 D_1^{-1} = (D_1^{-1} C^T D_0) A_0 (D_0 C D_1^{-1}).$$

Now, using the fact that the main diagonal entries of $A_1 \in M_m$ are all 1 for the first inequality and the monotonicity principle (Theorem 2.1) applied to $\lambda_n(A_0)K^T K \preceq K^T A_0 K$ with $K = D_0 C D_1^{-1}$ for the second, we have

$$m \ge \|A_1\| \ge \lambda_n(A_0)\|D_0 C D_1^{-1}\|^2.$$

Taking square roots and dividing by $\lambda_n^{\frac{1}{2}}(A_0)$ gives the asserted bound. □

If the matrix C is orthogonal, then $H_1 = U^T H_0 U$ implies that $H_0 = U H_1 U^T$ and so we have a companion bound stated in the next result.

COROLLARY 3.2. *Let $H_1 = U^T H_0 U \in M_n$ and let $H_i = D_i A_i D_i$, where the main diagonal entries of the A_i are 1 and the D_i are diagonal. Assume that U is orthogonal. Then*

$$\|D_0 U D_1^{-1}\| \le \lambda_1^{\frac{1}{2}}(A_1)\lambda_n^{-\frac{1}{2}}(A_0) \le \sqrt{n}\ \lambda_n^{-\frac{1}{2}}(A_0), \tag{3.28}$$

$$\|D_0^{-1} U D_1\| \le \lambda_1^{\frac{1}{2}}(A_0)\lambda_n^{-\frac{1}{2}}(A_1) \le \sqrt{n}\ \lambda_n^{-\frac{1}{2}}(A_1). \tag{3.29}$$

This says that if an orthogonal matrix U transforms H_0 into H_1 and $\lambda_n(D_i^{-1} H_i D_i^{-1}) = \lambda_n(A_i), i = 0, 1$ are not too small, then U has a graded structure.

In the special case that U is the matrix of eigenvectors of $H = DAD$, then $A_1 = I$ and we obtain

$$\|DU\Lambda^{-\frac{1}{2}}\| \le \lambda_n^{-\frac{1}{2}}(A), \quad \|D^{-1}U\Lambda^{\frac{1}{2}}\| \le \lambda_1^{\frac{1}{2}}(A) \le \sqrt{n}. \tag{3.30}$$

It is useful to have bounds on the individual entries of U and we state a variety of such bounds in (3.31)–(3.33) but note that they are actually weaker than the normwise bounds in (3.30). The bounds (3.31)–(3.33) are stronger than those in [2, Proposition 2.8] and [1, Proposition 6], which have a factor $\kappa^{\frac{3}{2}}(A)$ rather than $\kappa^{\frac{1}{2}}(A)$ on the right-hand side. The result in [1] is, however, applicable to scaled diagonally dominant symmetric matrices while our result is only for positive definite matrices.

COROLLARY 3.3. *Let $H = DAD \in M_n$ be positive definite and assume that D is diagonal and that the main diagonal entries of A are 1. Let U be an orthogonal matrix such that $\Lambda = U^T H U$ is diagonal with diagonal entries λ_i. Then*

$$|u_{ij}| \le \min\left\{\lambda_n^{-\frac{1}{2}}(A)\sqrt{\frac{\lambda_j}{h_{ii}}}, \lambda_1^{1/2}(A)\sqrt{\frac{h_{ii}}{\lambda_j}}\right\}, \tag{3.31}$$

$$|u_{ij}| \le \kappa^{\frac{1}{2}}(A)\min\left\{\sqrt{\frac{\lambda_i}{\lambda_j}}, \sqrt{\frac{\lambda_j}{\lambda_i}}\right\}, \tag{3.32}$$

$$|u_{ij}| \le \kappa^{\frac{1}{2}}(A)\min\left\{\sqrt{\frac{h_{ii}}{h_{jj}}}, \sqrt{\frac{h_{jj}}{h_{ii}}}\right\} \tag{3.33}$$

and the first inequality is stronger than the second and third.

Proof. The fact that $|u_{ij}|$ is no larger than the first (second) quantity on the right-hand side of (3.31) follows from the first (second) inequality in (3.30). The remaining inequalities can be derived from (3.31) using the relations between the eigenvalues of H and its main diagonal entries in Corollary 2.5. This also shows that they are weaker than (3.31). □

Another way to state the bound in (3.31) is

$$|U| \le \min\{\lambda_n^{-\frac{1}{2}}(A)D^{-1}E\Lambda^{\frac{1}{2}}, \lambda_1^{\frac{1}{2}}(A)DE\Lambda^{-\frac{1}{2}}\}, \tag{3.34}$$

where the minimum is taken componentwise. Recall that E is the matrix of ones.

3.2. Applications of graded eigenvectors. Now we use the results in section 3.1 to give another proof of the fact that components of the eigenvectors of a graded positive definite matrix are determined to a high relative accuracy. We then show that $\mathrm{relgap}^*(\lambda(H), i)$ is a good measure of the distance of a graded matrix from the nearest matrix with a multiple ith eigenvalue, where the distance is measured in a norm that respects that grading. Finally we show that Jacobi's method does indeed compute the eigenvectors to this accuracy (improving on [2, Theorem 3.4]).

We now combine Lemma 2.6 with the general technique used in section 2 to obtain a lemma that will be useful in proving componentwise bounds for eigenvectors and singular vectors.

LEMMA 3.4. *Let $\Lambda = \mathrm{diag}(\lambda)$ have main diagonal elements ordered in decreasing order and assume that $\lambda_{j+1} < \lambda_j < \lambda_{j-1}$. Let X be a symmetric matrix and let U be an orthogonal matrix. Let $H(\epsilon) = U\Lambda^{\frac{1}{2}}(I + \epsilon X)\Lambda^{\frac{1}{2}}U$. Let $u = Ue_j$ be an eigenvector of $H \equiv H(0)$ associated with λ_j. Let $\bar{u}$ be the upper bound on the jth eigenvector; that is,*

$$\bar{u}_i = \min\left\{\lambda_n^{-\frac{1}{2}}(A)\sqrt{\frac{\lambda_j}{h_{ii}}}, \lambda_1^{\frac{1}{2}}(A)\sqrt{\frac{h_{ii}}{\lambda_j}}\right\}.$$

Then, for ϵ sufficiently small, $\lambda_j(\epsilon) = \lambda_j(H(\epsilon))$ is simple, and one can choose $u(\epsilon)$ to be a unit eigenvector of $H(\epsilon)$ corresponding to $\lambda_j(\epsilon)$ such that

$$|u - u(\epsilon)| \le \frac{\sqrt{n-1}\,\|X\|\,\epsilon}{\mathrm{relgap}^*(\lambda, j)}\,\bar{u} \;+\; O(\epsilon^2). \tag{3.35}$$

Proof. Since U is the matrix of eigenvectors of H, the bound (3.31) gives

$$|u_{ik}| \le \bar{u}_{ik} \equiv \min\left\{\lambda_n^{-\frac{1}{2}}(A)\sqrt{\frac{\lambda_k}{h_{ii}}}, \lambda_1^{\frac{1}{2}}(A)\sqrt{\frac{h_{ii}}{\lambda_k}}\right\}. \tag{3.36}$$

Note that the vector $\bar{u}$ defined in the statement of the theorem is just the jth column of the matrix $\bar{U}$ just defined in (3.36). From Lemma 2.6 it follows that there is an eigenvector $\hat{u}(\epsilon)$ such that

$$|\hat{u}(\epsilon) - e_j| \le \epsilon r + O(\epsilon^2),$$

where r is the vector given by

$$r_k \equiv \begin{cases} 0, & k = j, \\ \left| \frac{x_{kj} \lambda_k^{\frac{1}{2}} \lambda_j^{\frac{1}{2}}}{\lambda_k - \lambda_j} \right|, & k \neq j, \end{cases}$$

where x_{ij} is the i,j element of X. Let $u(\epsilon) = U\hat{u}(\epsilon)$. So

$$|u - u(\epsilon)| = |U(e_j - \hat{u}(\epsilon))| \leq |U|\ |e_j - \hat{u}(\epsilon)| \leq \epsilon \bar{U}\ r + O(\epsilon^2),$$

and we must now bound $\bar{U}r$. The ith element of $\bar{U}r$ is

$$\begin{aligned}
\sum_{k=1}^{n} \bar{u}_{ik} r_k &\leq \sum_{k \neq j} \min \left\{ \lambda_n^{-\frac{1}{2}}(A) \sqrt{\frac{\lambda_k}{h_{ii}}}, \lambda_1^{\frac{1}{2}}(A) \sqrt{\frac{h_{ii}}{\lambda_k}} \right\} |\epsilon x_{kj}| \frac{\sqrt{\lambda_k \lambda_j}}{|\lambda_k - \lambda_j|} + O(\epsilon^2) \\
&= \epsilon \sum_{k \neq j} \min \left\{ \lambda_n^{-\frac{1}{2}}(A) \sqrt{\frac{\lambda_j}{h_{ii}}} \frac{\lambda_k}{|\lambda_k - \lambda_j|}, \lambda_1^{\frac{1}{2}}(A) \sqrt{\frac{h_{ii}}{\lambda_j}} \frac{\lambda_j}{|\lambda_k - \lambda_j|} \right\} |x_{kj}| + O(\epsilon^2) \\
&\leq \epsilon \sum_{k \neq j} \frac{\lambda_j + \lambda_k}{|\lambda_j - \lambda_k|} \min \left\{ \lambda_n^{-\frac{1}{2}}(A) \sqrt{\frac{\lambda_j}{h_{ii}}}, \lambda_1^{\frac{1}{2}}(A) \sqrt{\frac{h_{ii}}{\lambda_j}} \right\} |x_{kj}| + O(\epsilon^2) \\
&\leq \epsilon \cdot \mathrm{relgap}^{*-1}(\lambda, j) \sum_{k \neq j} \min \left\{ \lambda_n^{-\frac{1}{2}}(A) \sqrt{\frac{\lambda_j}{h_{ii}}}, \lambda_1^{\frac{1}{2}}(A) \sqrt{\frac{h_{ii}}{\lambda_j}} \right\} |x_{kj}| + O(\epsilon^2) \\
&= \epsilon \cdot \mathrm{relgap}^{*-1}(\lambda, j) \cdot \bar{u}_i \sum_{k \neq j} |x_{kj}| + O(\epsilon^2) \\
&\leq \epsilon \cdot \mathrm{relgap}^{*-1}(\lambda, j) \sqrt{n-1} \|X\|\ \bar{u}_i + O(\epsilon^2).
\end{aligned}$$

For the final equality note that the quantity

$$\min \left\{ \lambda_n^{-\frac{1}{2}}(A) \sqrt{\frac{\lambda_j}{h_{ii}}}, \lambda_1^{\frac{1}{2}}(A) \sqrt{\frac{h_{ii}}{\lambda_j}} \right\}$$

is now independent of k and is $\bar{u}_i$ as defined in the statement of this lemma. □

This result gives componentwise perturbation bounds for eigenvectors and singular vectors as simple corollaries.

THEOREM 3.5. *Let $H = DAD$ be positive definite and let $H(\epsilon) = D(A + \epsilon \Delta A)D$. Let $\eta = \|H^{-\frac{1}{2}}(\Delta H)H^{-\frac{1}{2}}\|$ and let $\lambda_j(\epsilon) = \lambda_j(H(\epsilon))$. Let $\lambda_j = \lambda_j(0)$ and assume that it is a simple eigenvalue of H. Let u be a corresponding unit eigenvector of H. Let $\bar{u}$ be the upper bound on the jth unit eigenvector; that is,*

$$\bar{u}_i = \min \left\{ \lambda_n^{-\frac{1}{2}}(A) \sqrt{\frac{\lambda_j}{h_{ii}}}, \lambda_1^{\frac{1}{2}}(A) \sqrt{\frac{h_{ii}}{\lambda_j}} \right\}.$$

Then, for sufficiently small ϵ, $\lambda_j(\epsilon)$ is simple and there is a unit eigenvector $u(\epsilon)$ of $H(\epsilon)$ corresponding to $\lambda_j(\epsilon)$ such that

$$|u - u(\epsilon)| \leq \frac{\sqrt{n-1}\ \epsilon \eta}{\mathrm{relgap}^*(\lambda, j)} \bar{u} + O(\epsilon^2).$$

Proof. Write

$$H(\epsilon) = U\Lambda^{\frac{1}{2}}(I + \epsilon\Delta)\Lambda^{\frac{1}{2}}U^T,$$

where U is the matrix of eigenvectors of H and $\Delta = \Lambda^{-\frac{1}{2}}U^T(\Delta H)U\Lambda^{-\frac{1}{2}}$. Lemma 2.2 implies that $\|\Delta\| = \eta$. The asserted bound now follows from Lemma 3.4. □

Lemma 3.4 also yields a componentwise bound on singular vectors.

THEOREM 3.6. *Let $G = BD \in M_{m,n}$ be of rank n and let $G(\epsilon) = (B + \epsilon\Delta B)D$. Let $\eta = \|(\Delta B)B^\dagger\|$. Let $\sigma = \sigma(G)$ and let $\sigma_j(\epsilon) = \sigma_i(G(\epsilon))$. Assume that σ_j is simple and that v is a corresponding unit right singular vector. Let $\bar{v}$ be the upper bound on the jth right unit singular vector; that is,*

$$\bar{v}_i = \min\left\{\sigma_n^{-1}(B)\frac{\sigma_j}{d_i}, \sigma_1(B)\frac{d_i}{\sigma_j}\right\}.$$

Then, for sufficiently small ϵ, $\sigma_j(\epsilon)$ is simple and there is a unit right singular vector $v(\epsilon)$ of $G(\epsilon)$ corresponding to $\sigma_j(\epsilon)$ such that

$$|v - v(\epsilon)| \leq \frac{2\sqrt{n-1}\,\epsilon\eta}{\text{relgap}^*(\sigma^2, j)}\bar{v} + O(\epsilon^2). \tag{3.37}$$

Proof. Let $G = U\Sigma V^T$ where $U \in M_{m,n}$ has orthonormal columns, $\Sigma \in M_n$ is positive diagonal, and $V \in M_n$ is orthogonal. We may write

$$G(\epsilon)^T G(\epsilon) = V\Sigma^T(I + \epsilon F)\Sigma V^T + O(\epsilon^2),$$

where $F = U^T B^{\dagger T}\Delta B^T U + U^T\Delta B B^\dagger U$, and $B^\dagger$ is the pseudoinverse of B. Note that $\|F\| \leq 2\eta$. Since the jth singular value of G is simple the corresponding singular vector is differentiable and so, in particular, $v(\epsilon)$, the jth singular vector of $G(\epsilon)$ (and eigenvector of $G(\epsilon)^T G(\epsilon)$), and $\hat{v}$, the jth eigenvector of $V\Sigma^T(I + \epsilon F)\Sigma V^T$, differ by at most $O(\epsilon^2)$. According to Lemma 3.4, we know that

$$|\hat{v} - v| \leq \frac{\sqrt{n-1}\,\|F\|\epsilon}{\text{relgap}^*(\sigma^2, j)}\,\bar{v} + O(\epsilon^2)$$

and hence the bound on $v(\epsilon)$. □

This improves [2, Proposition 2.20] in two ways. First, our upper bound $\bar{v}_j$ is smaller than that in [2] by a factor of about $\sigma_n^{-1}(B)$. Second, we have a factor $\sigma_n(B)$ in the denominator while in [2] there is a factor $\sigma_n^2(B)$, so overall our bound is smaller by a factor of about $\sigma_n^{-2}(B)$. The latter difference arises because in [2] the authors used the equivalent of Theorem 3.5 applied to $G^T G$, whereas we use Lemma 3.4.

The quantity $\text{relgap}^*(\sigma^2, j)$ can be hard to deal with when one perturbs G and hence also its singular values. It would be more convenient to have $\text{relgap}^*(\sigma, j)$ in the bound. It is easy to check that $\text{relgap}^*(\sigma, j) \leq \text{relgap}^*(\sigma^2, j)$. Thus (3.37) implies

$$|v - v(\epsilon)| \leq \frac{2\sqrt{n-1}\,\|\Delta B\|\,\epsilon}{\sigma_n(B)\text{relgap}^*(\sigma, j)}\bar{v} + O(\epsilon^2). \tag{3.38}$$

It is worth stating the stronger form of the inequality (3.37), as this is more natural when G is the Cholesky factor of a positive definite H (as is the case in [12]). In this case $\sigma^2(G) = \lambda(H)$.

Because we have no componentwise bound on the left singular vectors of $G = BD$ we cannot get a componentwise bound on the difference between the left singular vectors of BD and $(B + \Delta B)D$.

We now give a result on componentwise perturbations of singular vectors. Our bound is stronger than [2, Propositions 2.19 and 2.20] by a factor of about $\sigma_n^{-3}(B)$. (Our upper bound $\bar{v}$ is smaller than in [2] by a factor of $\sigma_n^{-2}(B)$, and the denominator here contains a factor $\sigma_n(B)$ while that in [2] contained $\sigma_n^2(B)$.) We could give an improved bound for eigenvectors also, but we restrict ourselves to the case of singular vectors because that is what is important when one uses one-sided Jacobi to compute eigenvectors of a positive definite matrix to high componentwise relative accuracy.

THEOREM 3.7. *Let $G = BD \in M_{m,n}$ have rank n and assume that $D = \operatorname{diag}(d_1, \ldots, d_n)$, where the d_i are positive and that B has unit columns. Choose $j \in \{1, \ldots, n\}$ and let v be a unit right singular vector corresponding to $\sigma_i(G)$. Let*

$$\bar{v} = \min\{\sigma_n^{-1}(B)\sigma_j(G)D^{-1}e, \sigma_1(B)\sigma_j^{-1}(G)De, \}.$$

Let $\Delta G = \Delta BD$ and set $\delta = \|\Delta B\|\sigma_n^{-1}(B)$ and assume that $\delta < \operatorname{relgap}^*(\sigma(G), j)$. Then $|v| \le \bar{v}$ and there is a vector $\hat{v}$ that is a right singular vector of $G + \Delta G$ such that

$$|v - \hat{v}| \le \frac{4\sqrt{n-1}\,(1 - \|\Delta B\|)^{-2}(1-\delta)^{-2}}{\sigma_n(B) \cdot (\operatorname{relgap}^*(\sigma(G), j) - \delta)}\, \|\Delta B\|\, \bar{v}. \tag{3.39}$$

Proof. The statement that $\bar{v}$ is an upper bound on v follows from (3.31). Let $G(t) = G + t\Delta G$. The condition $\delta < \operatorname{relgap}^*(\sigma(G), j)$ ensures that $\sigma_i(G(t))$ is simple for $t \in [0, 1]$ so there is a differentiable $v(t)$ that is a right singular vector of $G(t)$ such that $v(0) = v$, and from (3.38) we have the componentwise bound

$$\left|\frac{d}{dt}v(t)\right| \le \frac{2\sqrt{n-1}}{\sigma_n(B(t)) \cdot \operatorname{relgap}^*(\sigma(G(t)), j)}\, \|\Delta B\|\, \bar{v}(t), \tag{3.40}$$

where $G(t) = B(t)D(t)$ and $B(t)$ has unit columns and $D(t)$ is positive diagonal. So for a bound on $|v - \hat{v}| = |v(0) - v(1)|$ we need only bound each of the quantities that depend on t and then integrate the bound. Using the fact

$$1 - \delta \le \frac{\sigma_i(G(t))}{\sigma_i(G)} \le 1 + \delta,$$

one can show that for $t \in [0, 1]$

$$\operatorname{relgap}^*(\sigma(G(t)), j) \ge \operatorname{relgap}^*(\sigma(G), j) - \delta.$$

One can check that $B(t) = (B + t\Delta B)(DD(t)^{-1})$ so

$$\begin{aligned}\sigma_n(B(t)) &\ge \sigma_n(B + t\Delta B)\sigma_n(DD(t)^{-1}) \\ &\ge \sigma_n(B + t\Delta B)(1 - \|\Delta B\|) \ge \sigma_n(B)(1-\delta)(1 - \|\Delta B\|).\end{aligned}$$

The condition $\delta < \operatorname{relgap}^*(\sigma(G), j)$ implies $1 + \delta < 2$ because $\operatorname{relgap}^*(\sigma(G), j)$ is necessarily less than 1. We use this in the final inequality in the following display. Using (3.31) for the first inequality and bounds on $\sigma_j(G(t)), \sigma_n(B(t))$, and $d_i(t)$ for

the subsequent inequalities, we have

$$\begin{aligned}
|v_j(t)| &\le \min\left\{\sigma_n^{-1}(B(t))\frac{\sigma_j(G(t))}{d_i(t)},\ \frac{d_i(t)}{\sigma_j(G(t))}\right\}\\
&\le \min\left\{(1-\delta)^{-1}(1-\|\Delta B\|)^{-1}\sigma_n^{-1}(B(t))\frac{\sigma_j(G(t))(1+\delta)}{d_i(t)(1-\|\Delta B\|)}, \frac{d_i(t)(1+\|\Delta B\|)}{\sigma_j(G(t))(1-\delta)}\right\}\\
&\le (1-\delta)^{-1}(1-\|\Delta B\|)^{-1}(1+\delta)\min\left\{\sigma_n^{-1}(B)\frac{\sigma_j(G)}{d_i}, \frac{d_i}{\sigma_j(G)}\right\}\\
&\le 2(1-\delta)^{-1}(1-\|\Delta B\|)^{-1}\bar{v}_j.
\end{aligned}$$

Substituting these bounds into (3.40) gives

$$\left|\frac{d}{dt}v(t)\right| \le \frac{4\sqrt{n-1}\,(1-\|\Delta B\|)^{-2}(1-\delta)^{-2}}{\sigma_n(B)\cdot(\mathrm{relgap}^*(\sigma(G),j)-\delta)}\ \|\Delta B\|\ \bar{v},$$

which when integrated yields the asserted inequality. □

In right-handed Jacobi one computes the singular values of $G_0 \in M_n$ by generating a sequence $G_{i+1} = G_i J_i$, where J_i is an orthogonal matrix chosen to orthogonalize two columns of G_i. One stops when

$$|(G_M)_{i,\cdot}^T (G_M)_{j,\cdot}| \le \ tol \cdot \|(G_M)_{i,\cdot}\|\ \|(G_M)_{j,\cdot}\|, \quad i \ne j. \tag{3.41}$$

One can obtain the right singular vectors of G by accumulating the J_i. Demmel and Veselić show in [2, Theorem 3.4] that when implemented in finite precision arithmetic, this algorithm gives the individual components of the eigenvectors to a high accuracy relative to their upper bounds (actually this is for two-sided Jacobi, but the proof is essentially the same for one-sided Jacobi). However, their bound contains a factor for which they say "linear growth is far more likely than exponential growth." In the next result we show that the growth is indeed linear. One can prove an analogous result for two-sided Jacobi applied to a positive definite matrix.

Let us denote the product $J_i J_{i+1} \cdots J_k$ by $J_{i:k}$. The key idea that allows us to derive a growth factor that is linear in M rather than exponential in M is that we bound $J_{i:k}$ directly, rather than bound it by $|J_{i:k}| \le |J_i||J_{i+1}|\cdots|J_k|$ and then bounding each of the terms on the right-hand side. This idea has been used profitably in [11] also.

THEOREM 3.8. *Let $G_i = B_i D_i \in M_n, i = 0, 1, \ldots, M$, where B_i has unit columns and D_i is diagonal. Assume that*

$$G_{i+1} = (G_i + \Delta G_i)J_i,$$

where J_i is orthogonal and

$$\|\Delta G_i D_i^{-1}\| \le \eta.$$

Assume further that the columns of G_M are almost orthogonal in the sense that G_M satisfies (3.41) *with tolerance tol. Let*

$$\sigma_{\min} \equiv \min_{i=0,\ldots,M-1}\sigma_n(B_i)$$

and assume that $\delta \equiv M\sqrt{n}\eta\sigma_{\min}^{-1} < \mathrm{relgap}^*(\sigma(G), j)$. Let $\hat{u}$ be the computed column of $J_{0:M-1}$ corresponding to $\sigma_j(G)$. Then there is a unit right singular vector u_T of G corresponding to $\sigma_j(G)$ such that, to first order in η, ϵ, and tol,

$$|u - \hat{u}| \le \left[\frac{Mn^{3/2}}{\sigma_{\min}^{-2}}\epsilon + \frac{2\sqrt{n-1}(M\sigma_{\min}^{-1}\eta + n\cdot tol)}{\mathrm{relgap}^*(\sigma(G),j) - 2M\sigma_{\min}^{-1}\eta - n\cdot tol}\right]\bar{u}, \tag{3.42}$$

where

$$\bar{u} = \min\{\sigma_n^{-1}(B)\sigma_j(G)D^{-1}e, \sigma_1(B)\sigma_j^{-1}(G)De\}$$

is the upper bound on u from (3.31).

The bound (3.42) is a first-order bound. The proof below would also yield a bound that takes into account all the higher-order terms, but the resulting inequality would be much more complicated and probably not any more useful.

If the G_i and J_i arise from right-handed Jacobi applied to G in finite precision arithmetic with precision ϵ, then one can take $\eta = 12\epsilon$ [10, Theorem 4.2 and the ensuing discussion].

Let us compare our bound with

$$q(M,n)\frac{(tol+\epsilon)\cdot n}{\text{relgap}^*(\sigma(G),j)\ \sigma_{\min}^2}\ \bar{\bar{u}}, \tag{3.43}$$

which is essentially the bound on the computed right singular values from [2, Theorem 4.4] stated in our notation. Our bound is stronger in several respects. The term $q(M,n)$ is a growth factor that is exponential in M, while our bound is linear in M. As we have mentioned earlier, the upper bound vector $\bar{\bar{u}}$ in (3.43) is larger than $\bar{u}$ by a factor of about $\sigma_n^{-2}(B)$, which could be quite significant. Also, we have two terms, one in $\sigma_{\min}^{-2}$ and the other in $(\sigma_{\min}\cdot\text{relgap}^*(\sigma(G),j)^{-1}$; both these quantities are less than $(\sigma_{\min}^2\text{relgap}^*(\sigma(G),j)^{-1}$, which occurs in (3.43).

A weakness of both bounds is that they contain the factor $\sigma_{\min}^{-1}$ (defined in the statement of the theorem) rather than $\sigma_n^{-1}(B)$. It has been observed experimentally [2, section 7.4] that $\sigma_{\min}/\sigma_n(B)$ is generally 1 or close to 1, but no rigorous proof of this fact is known. Mascarenhas has shown that the ratio can be as large as $n/4$ [8].

One can also see that for a given ϵ we can take *tol*, the stopping tolerance, as large as $\epsilon\sigma_n^{-1}(B)$ without significantly increasing the right-hand side of (3.42). Typically, it is suggested that one take *tol* to be a modest multiple of ϵ when one wants to compute the eigenvectors or eigenvalues to high relative accuracy [2]. Thus this larger value of *tol* may be useful in practice to save a little computation through earlier termination.

The rest of the paper is devoted to the rather lengthy proof of this theorem.

Proof. The outline of the proof is as follows. First we will bound $|u-\hat{u}|$, where u is the value of the jth column of $J_{0:M-1}$ computed in exact arithmetic. Next, we will bound $|u-u_T|$, where u_T (T is for true) is an exact singular vector of G associated with $\sigma_j(G)$. The inequality (3.42) follows by combining these two bounds. Throughout we will use the facts that $\sigma_j(G_M) = \sigma_j(G)(1+\alpha_j)$, where $|\alpha_j| \le M\eta\sigma_{\min}^{-1}$ and $|\sigma_n(B_M)-1| \le n\cdot tol$ and drop second-order terms.

Now consider $|u-\hat{u}|$. This bound depends only on the scaling of the $J_{i:k}$ and is independent of $\text{relgap}^*(\sigma(G),j)$. If $X, Y \in M_n$ are multiplied in floating point arithmetic with precision ϵ, the result is $XY+\Delta$, where

$$|\Delta| \le 2n\epsilon|X||Y|.$$

Using this fact, one can show by induction that

$$J_{0:M-1} - \hat{J}_{0:M-1} \le 2n\epsilon\sum_{i=1}^{M-1} E_i J_{i+1:M-1} + O(\epsilon^2), \tag{3.44}$$

where $|E_i| \le |J_{0:i-1}||J_i|$. Here E_i is the error in multiplying $J_{0:i-1}$ and J_i, and $E_i J_{i+1:M-1}$ is the first-order effect of this error in the computed value of $J_{0:M}$. Taking

absolute values in (3.44) gives the componentwise error bound

$$|J_{0:M-1} - \hat{J}_{0:M-1}| \leq 2n\epsilon \sum_{i=1}^{M-1} |J_{0:i-1}||J_i||J_{i+1:M-1}| + O(\epsilon^2). \tag{3.45}$$

Now

$$\begin{aligned} D_0|J_{0:i-1}||J_i||J_{i+1:M-1}|D_M^{-1} &= (D_0|J_{0:i-1}|D_i^{-1})(D_i|J_i|D_{i+1}^{-1})(D_{i+1}|J_{i+1:M-1}|D_M^{-1}) \\ &\leq (D_0|J_{0:i-1}|D_i^{-1})(D_i|J_i|D_{i+1}^{-1}) \cdot \sigma_n^{-1}(B_{i+1})E \\ &\leq \|D_0|J_{0:i-1}|D_i^{-1}\| \, \|D_i|J_i|D_{i+1}^{-1}\| \, \sqrt{n}\sigma_n^{-1}(B_{i+1})E \\ &\leq \sqrt{n}\sigma_n^{-1}(B_0) \cdot \sqrt{n}\sigma_n^{-1}(B_i)\sqrt{n}\sigma_n^{-1}(B_{i+1})E \\ &\leq n^{3/2}\sigma_n^{-1}(B_0)\sigma_{\min}^{-2}E. \end{aligned}$$

Recall that E denotes the matrix of ones. We have used the first term in (3.34) and the fact that, up to first order, $J_{i+1:M-1}$ diagonalizes $G_{i+1}^T G_{i+1}$ for the first inequality and we have used (3.28) twice for the third inequality. Since G_M has orthogonal columns up to $O(tol)$ and the singular values of G_M are the same as those of G to $O(\eta)$, it follows that $D_M = \Sigma$ at least to first order. So, multiplying by D_0 and Σ^{-1}, we have, to first order,

$$D_0|J_{0:M-1} - \hat{J}_{0:M-1}|\Sigma^{-1} \leq Mn^{3/2}\sigma_n^{-1}(B_0)\sigma_{\min}^{-2}\epsilon E.$$

In the same way, we obtain the first-order bound

$$D_0^{-1}|J_{0:M-1} - \hat{J}_{0:M-1}|\Sigma \leq Mn^{3/2}\sigma_{\min}^{-2}\epsilon E.$$

These two bounds can be combined to give

$$|J_{0:M-1} - \hat{J}_{0:M-1}| \leq Mn^{3/2}\sigma_{\min}^{-2} \cdot \epsilon \cdot \min\{\sigma_n^{-1}(B_0)D^{-1}E\Sigma, DE\Sigma^{-1}\},$$

where the minimum is taken componentwise. (Note that $D_0 = D$.) The jth column of this is the inequality we desire:

$$|u - \hat{u}| \leq Mn^{3/2}\sigma_{\min}^{-2}\epsilon \min\{\sigma_n^{-1}(B_0)\sigma_j(G)D^{-1}e, \sigma_j^{-1}(G)De\} = Mn^{3/2}\sigma_{\min}^{-2}\epsilon\bar{u}.$$

This completes the first step.

Now let us bound the error between u and a singular vector of G. If the columns of G_M were orthogonal, then e_j, in particular, would be a right singular vector associated with $\sigma_j(G_M)$. If, in addition, all the ΔG_i were 0, then $G_M = G_0 J_{0:M-1}$ and so $u = J_{0:M-1}e_j$ would be a right singular vector associated with $\sigma_j(G)$. Neither of these hypotheses is true, though in each case they are "almost true" and so u is close to being a singular vector of G_0. We now bound the difference.

First we will consider the fact that the columns of G_M are not exactly orthogonal. Write

$$G_M^T G_M = D_M(I + A)D_M = D_M(I + A)^{\frac{1}{2}}(I + A)^{\frac{1}{2}}D_M. \tag{3.46}$$

Then each entry of A is at most tol in absolute value and so $\|A\| \leq n \cdot tol$. The equation (3.46) implies that there is an orthogonal matrix Q such that $QG_M = (I + A)^{\frac{1}{2}}D_M$. One can check that

$$\|I - (I + A)^{\frac{1}{2}}\| = \max_i\{|1 - \sqrt{1 + \lambda_i(A)}|\} \leq \|A\| \leq n \cdot tol;$$

we will use this bound later.

Now consider the effect of the ΔG_i. It is easy to check by induction, for example, that

$$G_M = G_0 J_{0:M-1} + \Delta G, \tag{3.47}$$

where

$$\Delta G = \sum_{i=0}^{M-1} \Delta G_i J_{i:M-1}.$$

Now, using the assumption $\|\Delta G_i D_i^{-1}\| \le \eta$ for the first inequality and (3.28) for the second, we have

$$\begin{aligned}\|\Delta G_i J_{i:M-1} D_M^{-1}\| &= \|\Delta G_i D_i^{-1} D_i J_{i:M-1} D_M^{-1}\| \\ &\le \|\Delta G_i D_i^{-1}\| \, \|D_i J_{i:M-1} D_M^{-1}\| \\ &\le \eta \sigma_{\min}^{-1}.\end{aligned}$$

Together with (3.47) this yields

$$\|\Delta G D_M^{-1}\| \le M \sigma_{\min} \eta.$$

Now we will combine these two results to show that $G_M + \Delta G$ has a right singular vector close to e_j and hence that $G_0 = (G_M + \Delta G) J_{0:M-1}^T$ has a right singular vector close to $u = J_{0:M-1} e_j$. The right singular vectors of $G_M + \Delta G$ are the same as those of $Q(G_M + \Delta G)$ where Q is the orthogonal matrix introduced after equation (3.46). Also,

$$\begin{aligned}Q(G_M + \Delta G) &= (I + A)^{\frac{1}{2}} D_M + Q \Delta G \\ &= D_M + [(I + A)^{\frac{1}{2}} - I + Q \Delta G D_M^{-1}] D_M.\end{aligned}$$

The jth right singular vector of D_M is e_j and

$$\begin{aligned}\|(I + A)^{\frac{1}{2}} - I + Q^T \Delta G D_M^{-1}\| &\le \|(I + A)^{\frac{1}{2}} - I\| + \|Q^T \Delta G D_M^{-1}\| \\ &\le n \cdot tol + \|\Delta G D_M^{-1}\| \\ &\le n \cdot tol + M \sigma_{\min} \eta \\ &\equiv \tau.\end{aligned}$$

So by Theorem 3.7 there is a right singular vector v of $G + \Delta G$ corresponding to its jth singular value such that

$$|e_j - v| \le \frac{\sqrt{n-1}(1-\tau)^{-2}(1 - \sigma_n^{-1}(B_M)\tau)^{-2}}{\sigma_n(B_M) \cdot [\text{relgap}^*(\sigma(G_M), j) - \tau \sigma_n(B_M)]} \, \tau \bar{v}, \tag{3.48}$$

where

$$\bar{v} = \min\{\sigma_n^{-1}(B_M)\sigma_j(G_M) D_M^{-1} e, \sigma_1(B_M)\sigma_j^{-1}(G_M) D_m e\}.$$

Now let us drop the second-order terms in (3.48) and $\bar{v}$. The term τ is $O(\eta) + O(tol)\sigma_1(B_M)$, so we may drop all first-order terms in $\bar{v}$ and in the ratio in (3.48). In particular, we may replace $\sigma_n(B_M)$, $1 - \tau$, and $1 - \sigma_n^{-1}(B_M)\tau$ all by 1. We do not

assume that $\mathrm{relgap}^*(\sigma(G_M), j)$ is large with respect to η and τ so we cannot replace $\mathrm{relgap}^*(\sigma(G_M), j) - \tau$ by $\mathrm{relgap}^*(\sigma(G), j) - \tau$. However, as was shown at the end of the introduction, we have

$$\mathrm{relgap}^*(\sigma(G_M), j) \geq \mathrm{relgap}^*(\sigma(G), j) - M\eta\sigma_{\min}^{-1},$$

and hence

$$\mathrm{relgap}^*(\sigma(G_M), j) - \tau \geq \mathrm{relgap}^*(\sigma(G), j) - \gamma,$$

where $\gamma = 2M\sigma_{\min}\eta + n \cdot tol$. With these substitutions we obtain the bound that is equivalent to (3.48) up to first order

$$|e_j - v| \leq \frac{2\sqrt{n-1}(M\eta\sigma_{\min}^{-1} + n \cdot tol)}{\mathrm{relgap}^*(\sigma(G_M), j) - \gamma}\bar{\bar{v}}, \tag{3.49}$$

where

$$\bar{\bar{v}} = \min\{\sigma_j(G)D_M^{-1}e, \sigma_j^{-1}(G)D_M e\}.$$

For convenience, let the coefficient of $\bar{\bar{v}}$ in (3.49) be denoted by c.

Let $u_T = J_{0:M-1}v$. By construction it is a right singular vector of G_0 corresponding to $\sigma_j(G_0)$. Now we can complete the proof by bounding $|u - u_T|$.

$$|u - u_T| = |J_{0:M-1}e_j - J_{0:M-1}v| \leq |J_{0:M-1}|\ |e_j - v| \leq c \cdot |J_{0:M-1}|\bar{\bar{v}}.$$

So

$$\begin{aligned} D|u - u_T| &\leq c \cdot D|J_{0:M-1}|\bar{\bar{v}} \\ &\leq c \cdot D|J_{0:M-1}|D_M^{-1}D_M\bar{\bar{v}} \\ &\leq c \cdot [D|J_{0:M-1}|D_M^{-1}]\sigma_j(G)e \\ &\leq c \cdot [\sigma_1(B_M)\sigma_n^{-1}(B)E]\sigma_j(G)e \\ &\leq c \cdot n\sigma_n^{-1}(B)\sigma_j(G)e. \end{aligned}$$

We have used a slight generalization of (3.31) for the penultimate inequality and have dropped second-order terms in the last inequality. Similarly,

$$D^{-1}|u - u_T| \leq c \cdot D^{-1}|J_{0:M-1}|\bar{\bar{v}} \leq c \cdot n \cdot \sigma_j^{-1}(G)e,$$

and so

$$|u - u_T| \leq c \cdot n \cdot \min\{\sigma_n^{-1}(B)\sigma_j(G)D^{-1}e, \sigma_j^{-1}(B)De\} = cn\bar{u}.$$

Now combine the bound on $|u - u_T|$ and $|u - \hat{u}|$. □

Acknowledgment. I thank the referee whose detailed comments, including the observation that grading is not necessary for relative perturbation bounds, have greatly improved the presentation of the results in this paper.

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] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.
[3] S. Eisenstat, Personal communication, 1996.
[4] S. C. Eisenstat and I. C. F. Ipsen, *Relative perturbation techniques for singular value problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1972–1988.
[5] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[6] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.
[7] R.-C. Li, *Relative Perturbation Theory:* (I) *Eigenvalue and Singular Value Variations*, LAPACK Working Note 84, SIAM J. Matrix Anal., to appear.
[8] W. Mascarenhas, *A note on Jacobi being more accurate then QR*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 215–218.
[9] R. Mathias, *A bound for the matrix square root with application to eigenvector perturbation*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 861–867.
[10] R. Mathias, *Accurate eigensystem computations by Jacobi methods*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 977–1003.
[11] R. Mathias, *Instability of parallel prefix matrix multiplication*, SIAM J. Sci. Statist. Comput., 16 (1995), pp. 956–973.
[12] R. Mathias, *Fast accurate eigenvalue computations for graded positive definite matrices,* Numer. Math., 74 (1996), pp. 85–103.
[13] R. Mathias and K. Veselić, *A relative perturbation bound for positive definite matrices*, Linear Algebra Appl., to appear.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 4, pp. 981–999, October 1997

012

NUMERICAL CONDITION OF DISCRETE WAVELET TRANSFORMS*

RADKA TURCAJOVÁ†

Abstract. The recursive algorithm of a (fast) discrete wavelet transform, as well as its generalizations, can be described as repeated applications of block-Toeplitz operators or, in the case of periodized wavelets, multiplications by block circulant matrices. Singular values of a block circulant matrix are the singular values of some matrix trigonometric series evaluated at certain points. The norm of a block-Toeplitz operator is then the essential supremum of the largest singular value curve of this series. For all reasonable wavelets, the condition number of a block-Toeplitz operator thus is the lowest upper bound for the condition of corresponding block circulant matrices of all possible sizes. In the last section, these results are used to study conditioning of biorthogonal wavelets based on B-splines.

Key words. biorthogonal wavelets, block circulant matrices, block-Toeplitz operators, numerical condition, translational bases

AMS subject classifications. 15A12, 15A60, 42A45, 42C15

PII. S0895479894278319

1. Introduction.

Orthogonality is a very strong property. It might exclude other useful properties like, for example, symmetry in the case of compactly supported wavelets [6, 7]. Consequently, in many applications biorthogonal wavelets have been used rather than the orthogonal ones. Stability of such bases has been studied and conditions for Riesz bounds to be finite were established [2, 3, 4, 5]. However, when dealing with applications, one would like to have some *quantitative* information about sensitivity to such things like noise in the data or quantization. Some relevant estimates can be found in the engineering literature on multirate filter banks, where noise is modeled as a random process and its transmission through the system is studied; see, e.g., [12]. However, most of these results concern particular designs and implementations. Here we will use an alternative approach—we will look at discrete wavelet transforms from the point of view of linear algebra.

For example, let us consider the process of image compression using wavelets (see, e.g., [1, 11, 14]). The algorithm has three steps. First, the discrete wavelet transform is applied to the image, then the resulting data is quantized, and finally it is coded in some efficient way. The purpose of the transform is to increase the compressibility of the data and to restructure the data so that, after decompression, the error caused by quantizing is less disturbing for a human viewer than if the image were quantized directly without a transform. The encoded image can be manipulated in different ways (e.g., transmitted over networks) which can cause further distortions. To decompress the image we just need to decode the data and to apply the inverse transform. Let us denote the error vector that is added to the transformed data $\boldsymbol{y}$ before the reconstruction by $\boldsymbol{u}$ and let us suppose that we know the magnitude of the relative error, $\|\boldsymbol{u}\|/\|\boldsymbol{y}\| = \alpha$. Then, if $\boldsymbol{x}$ denotes the original image, the relative error

*Received by the editors November 21, 1994; accepted for publication (in revised form) by G. Cybenko October 30, 1996.

http://www.siam.org/journals/simax/18-4/27831.html

†National Institute of Standards and Technology, Building 820, Room 365, Gaithersburg, MD 20899-0001 (radka@snad.ncsl.nist.gov). This work was supported by the Flinders University of South Australia, the Cooperative Research Centre for Sensor Signal and Information Processing, Adelaide, and the Australian Government.

in the reconstructed image is

$$\frac{\|T^{-1}\boldsymbol{u}\|}{\|\boldsymbol{x}\|} \le \frac{\|T^{-1}\|\,\|\boldsymbol{u}\|}{\|\boldsymbol{x}\|} = \frac{\|T^{-1}\|\,\alpha\|\boldsymbol{y}\|}{\|\boldsymbol{x}\|} \le \frac{\|T^{-1}\|\,\alpha\,\|T\|\|\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \|T^{-1}\|\,\|T\|\,\alpha.$$

If no further assumptions are imposed on the image and type of the error, this estimate is the best possible. Also, in other applications, the sensitivity to errors can be shown to be naturally related to the condition number of the transform matrix with respect to solving a system of linear equations,

$$\operatorname{cond}(T) = \|T\|\,\|T^{-1}\|. \tag{1.1}$$

Condition number depends on the norm. For finite matrices we will use here the matrix 2-norm, which is induced by the Euclidean vector norm. When necessary, we will use subscript 2 to emphasize that we deal with these norms. We will speak also about the condition number of an operator $l^2(\mathbb{Z}) \longrightarrow l^2(\mathbb{Z})$. We define it also by (1.1); the norm is the operator norm induced by the norm of $l^2(\mathbb{Z})$.

Due to the translational character of wavelet bases, matrices and operators involved happen to have a characteristic structure—they are block circulant and block-Toeplitz, respectively. This structure can be employed when the condition numbers are computed; Fourier techniques can be used to transform them to a block diagonal form. This then leads to studying the (pointwise) singular values of certain trigonometric matrix series. In section 3 we study the finite case. The singular values of a block circulant matrix are shown to be the singular values of small matrices arising from the "block discrete Fourier transform" of the first block row of the block circulant matrix. In section 4 we generalize this result for block-Toeplitz operators $l^2(\mathbb{Z}) \longrightarrow l^2(\mathbb{Z})$. The situation is rather more complicated there, because the Fourier transform maps the discrete space $l^2(\mathbb{Z})$ onto the functional space $L^2([0, 2\pi))$. As the main result we show there that

$$\|C(A)\|_{\mathcal{B}(l^2(\mathbb{Z}))} = \operatorname*{ess\,sup}_{\xi\in[0,2\pi)} \sigma_{\max}(\mathcal{A}(\xi)),$$

where $C(A)$ is the block-Toeplitz operator the infinite matrix of which is generated by the strip

$$A = (\,\cdots \quad A_{-1} \quad A_0 \quad A_1 \quad A_2 \quad \cdots\,)$$

(A_j, $j \in \mathbb{Z}$, being square blocks) and

$$\mathcal{A}(\xi) = \sum_{k\in\mathbb{Z}} A_k e^{ik\xi}.$$

The proof is based on the pointwise singular decomposition of $\mathcal{A}$; some difficulties arising from the fact that we have to ensure that the singular vector we want to construct has square integrable components must be overcome on the way. For reasonable wavelets, the curves of singular values of $\mathcal{A}$ have some smoothness, and essential supremum and infimum become supremum and infimum or even maximum and minimum. The condition number of $C(A)$ is then the lowest upper bound on the condition of periodized wavelet transforms for all possible lengths of data. We also describe how some particular properties of the wavelets imply a certain structure of the singular values. These observations can be used to further improve the efficiency of computing the condition numbers.

In the last section of this paper, we apply this technique to study conditioning of biorthogonal B-spline wavelets constructed by Cohen, Daubechies, and Feauveau [5], which are probably the most often applied biorthogonal wavelets today. We show there that the condition number increases exponentially with the order of the spline. Conditioning can be significantly improved by suitable scaling of the wavelet functions, but, even for the optimal scaling, the growth has exponential character.

After finishing the first version of this paper, we became familiar with related works by Keinert [8] and Strang [9]. While Strang's work concerns mostly Riesz bounds for subspaces in a multiresolution analysis and wavelet decomposition, Keinert concentrates on conditioning of finitely sized transforms and asymptotic estimates for deep recursive transforms. He also presents a number of numerical experiments that show how these estimates are realistic when some specific types of introduced errors are considered (e.g., white noise). In this revised version we have tried to emphasize results that are complementary to those of Keinert and Strang.

2. Translational and wavelet bases and the operators of the change of a basis. Let us consider some translation-invariant subspace of $L^2(\mathbb{R})$ with a translational Riesz basis $\{u_k(x-hn), k=1,\dots,r, n\in\mathbb{Z}\}$ generated by some r-tuple of functions u_k, $k=1,\dots,r$, h being the translation step. Let this subspace have another, similar, basis $\{v_k(x-hn), k=1,\dots,r, n\in\mathbb{Z}\}$. Each of the functions v_k, $k=1,\dots,r$, can be expressed in terms of the first basis; there exist sequences $\{a_n^{(k,l)}\}_{n\in\mathbb{Z}}\in l^2(\mathbb{Z})$ such that

$$v_k(x)=\sum_{n\in\mathbb{Z}}\sum_{l=1}^{r} a_n^{(k,l)}u_l(x-hn).$$

Let us form from these coefficients $r\times r$ matrices A_n, $n\in\mathbb{Z}$; $a_n^{(k,l)}$ will be the element of A_n in the kth row and lth column. We denote by A the infinite strip of concatenated matrices A_n, $n\in\mathbb{Z}$,

$$A=(\cdots\quad A_{-1}\quad A_0\quad A_1\quad A_2\quad\cdots),$$

and we define $C(A)$ to be an infinite *block-Toeplitz* matrix

$$C(A)=\begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & A_0 & A_1 & A_2 & A_3\cdots & \\ \cdots & A_{-1} & A_0 & A_1 & A_2\cdots & \\ \cdots & A_{-2} & A_{-1} & A_0 & A_1\cdots & \\ \cdots & A_{-3} & A_{-2} & A_{-1} & A_0\cdots & \\ & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

We will also denote by $C(A)$ an operator $l^2(\mathbb{Z})\longrightarrow l^2(\mathbb{Z})$ that can be represented by such a matrix. If

$$\sum_{n\in\mathbb{Z}}\sum_{k=1}^{r}\overline{\alpha}_{nr+k-1}u_k(x-hn)=\sum_{n\in\mathbb{Z}}\sum_{k=1}^{r}\overline{\beta}_{nr+k-1}v_k(x-hn)$$

for some $l^2(\mathbb{Z})$ sequences $\{\alpha_n\}_{n\in\mathbb{Z}}$ and $\{\beta_n\}_{n\in\mathbb{Z}}$, then $C(A)^*$ maps $\{\beta_n\}_{n\in\mathbb{Z}}$ to $\{\alpha_n\}_{n\in\mathbb{Z}}$; that is, it is the operator of the change of a basis.

Because of practical reasons (handling of finite data), periodized bases are often used in the wavelet context. If, for some integer N, we denote

$$u_k^{per}(x) = \sum_{l\in\mathbb{Z}} u_k(x - Nhl), \qquad v_k^{per}(x) = \sum_{l\in\mathbb{Z}} v_k(x - Nhl),$$

then $\{u_k^{per}(x-n), k = 1,\dots,r, n = 0,\dots,N-1\}$ and $\{v_k^{per}(x-n), k = 1,\dots,r, n = 0,\dots,N-1\}$ are bases for some subspace of $L^2([0, Nh))$, and the operator of the change of basis from the latter to the former can be represented by a *block circulant* matrix

$$C_N(A) = \begin{pmatrix} S_0 & S_1 & S_2 & \cdots & S_{N-1} \\ S_{N-1} & S_0 & S_1 & \cdots & S_{N-2} \\ S_{N-2} & S_{N-1} & S_0 & \cdots & S_{N-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_1 & S_2 & S_3 & \cdots & S_0 \end{pmatrix},$$

where

$$S_k = \sum_{l\in\mathbb{Z}} A_{k+lN}.$$

A multiresolution analysis is a sequence of embedded subspaces of $L^2(\mathbb{R})$ generated by the translations of an appropriately dilated scaling function. In particular,

$$V_j = \overline{\mathrm{Span}\{2^{j/2}\varphi(2^j x - k), k \in \mathbb{Z}\}}.$$

There are wavelet subspaces generated by a wavelet function,

$$W_j = \overline{\mathrm{Span}\{2^{j/2}\psi(2^j x - k), k \in \mathbb{Z}\}},$$

and these subspaces satisfy

$$V_j = V_{j-1} \oplus W_{j-1}.$$

The scaling and wavelet functions thus have to conform to the two-scale relations that are usually written as

$$\varphi(x) = \sqrt{2}\sum_{k\in\mathbb{Z}} h_k \varphi(2x - k), \qquad \psi(x) = \sqrt{2}\sum_{k\in\mathbb{Z}} g_k \varphi(2x - k). \tag{2.1}$$

In the (fast) discrete wavelet transform, we perform recursively the change of basis from $\{2^{j/2}\varphi(2^j x - k), k \in \mathbb{Z}\}$ to $\{2^{(j-1)/2}\varphi(2^{j-1}x - k), k \in \mathbb{Z}\} \cup \{2^{(j-1)/2}\psi(2^{j-1}x - k), k \in \mathbb{Z}\}$, $j = J, J-1, J-2, \dots$. We can consider both bases to be generated by *two* functions, the former by $u_1 = 2^{j/2}\varphi(2^j x)$ and $u_2 = 2^{j/2}\varphi(2^j x - 1)$, the latter by $v_1 = 2^{(j-1)/2}\varphi(2^{j-1}x)$ and $v_2 = 2^{(j-1)/2}\psi(2^{j-1}x)$. The translation step h is $2^{-(j-1)}$ here. The recursive inverse transform can thus be associated with repeated applications of $C(A)^*$, where

$$A_n = \begin{pmatrix} h_{2n} & h_{2n+1} \\ g_{2n} & g_{2n+1} \end{pmatrix};$$

that is,

$$A = \begin{pmatrix} \cdots & h_{-2} & h_{-1} & h_0 & h_1 & h_2 & h_3 & \cdots \\ \cdots & g_{-2} & g_{-1} & g_0 & g_1 & g_2 & g_3 & \cdots \end{pmatrix}.$$

In fact, $C_(A)^* = C(\tilde{A})^{-1}$ for some $\tilde{A}$, and the recursive transform itself can be seen as repetitive applications of a block-Toeplitz operator. As in the case of A,

$$\tilde{A} = \begin{pmatrix} \cdots & \tilde{h}_{-2} & \tilde{h}_{-1} & \tilde{h}_0 & \tilde{h}_1 & \tilde{h}_2 & \tilde{h}_3 & \cdots \\ \cdots & \tilde{g}_{-2} & \tilde{g}_{-1} & \tilde{g}_0 & \tilde{g}_1 & \tilde{g}_2 & \tilde{g}_3 & \cdots \end{pmatrix};$$

sequences $\{\tilde{h}_n\}_{n\in\mathbb{Z}}$ and $\{\tilde{g}_n\}_{n\in\mathbb{Z}}$ determine the biorthogonal counterparts of the scaling and wavelet functions, $\tilde{\varphi}$ and $\tilde{\psi}$, by relations analogous to (2.1).

Although the conditioning of this basic step of the recursive transform is crucial, we also want to study how the error accumulates in the recursive transform. Since all the bases involved have translational character, we can use the same approach as for one step for the transform of any finite depth; we can always find a common translation step. For example, let us consider two steps of recursion. We perform, in fact, the change of basis in V_j from $\{2^{j/2}\varphi(2^j x - k), k \in \mathbb{Z}\}$ to $\{2^{(j-2)/2}\varphi(2^{j-2}x - k), k \in \mathbb{Z}\} \cup \{2^{(j-2)/2}\psi(2^{j-2}x - k), k \in \mathbb{Z}\} \cup \{2^{(j-1)/2}\psi(2^{j-1}x - k), k \in \mathbb{Z}\}$. All these bases can be considered to be translational bases with translation step $h = 2^{-(j-2)}$ generated by four functions. We have

$$u_n = 2^{j/2}\varphi(2^j x - (n-1)), \qquad n = 1, \ldots, 4,$$

and

$$\begin{aligned} v_1 &= 2^{(j-2)/2}\varphi(2^{j-2}x), \qquad & v_3 &= 2^{(j-1)/2}\psi(2^{j-1}x), \\ v_2 &= 2^{(j-2)/2}\psi(2^{j-2}x), \qquad & v_4 &= 2^{(j-1)/2}\psi(2^{j-1}x - 1). \end{aligned}$$

The infinite strip A will thus have four rows; the entries can be easily found by recursive applications of (2.1). In particular, if we denote the sequences that form rows of A by $\{b_n^{(s)}\}_{n\in\mathbb{Z}}$, $s = 1, \ldots, 4$, $b_0^{(s)} = a_0^{(s,1)}$, we have

$$\begin{aligned} b_n^{(1)} &= \sum_{k\in\mathbb{Z}} h_k h_{n-2k}, \qquad & b_n^{(3)} &= g_n, \\ b_n^{(2)} &= \sum_{k\in\mathbb{Z}} g_k h_{n-2k}, \qquad & b_n^{(4)} &= g_{n-2}. \end{aligned}$$

An analogous approach can be used for generalizations of classical wavelet transforms like those based on more than one scaling and wavelet function and general integer dilation parameter $m \geq 2$ (multiwavelets, higher multiplicity wavelets) or nonstationary wavelets, where different block-Toeplitz operators applied in the recursive algorithm. Also, wavelet packets transforms, where wavelet spaces are also further decomposed, can be described in a similar way.

3. Numerical condition of block circulant matrices. Any circulant matrix is unitarily similar to a diagonal matrix. This matrix has (up to scale) the discrete Fourier transform of the first row of the original matrix on the diagonal, and the similarity matrix is the matrix of the discrete Fourier transform itself. This fact can be generalized for block circulant matrices as follows.

THEOREM 3.1. *Each block circulant matrix is unitarily similar to a block diagonal matrix. In particular, $C_N(A)$ is similar to a matrix with diagonal blocks equal to $\mathcal{A}(2\pi i n/N)$, $n = 0, \ldots, N-1$, where $\mathcal{A}(\xi) = \sum_{k\in\mathbb{Z}} A_k e^{ik\xi}$.*

Proof. Let ω_N be the primitive Nth root of unity, $\omega_N = e^{2\pi i/N}$, and let us first create the matrix of the "block discrete Fourier transform";

$$\Omega_{r,N} = \frac{1}{\sqrt{N}} \begin{pmatrix} \omega_N^0 I & \omega_N^0 I & \omega_N^0 I & \cdots & \omega_N^0 I \\ \omega_N^0 I & \omega_N^{-1} I & \omega_N^{-2} I & \cdots & \omega_N^{-(N-1)} I \\ \omega_N^0 I & \omega_N^{-2} I & \omega_N^{-4} I & \cdots & \omega_N^{-2(N-1)} I \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \omega_N^0 I & \omega_N^{-(N-1)} I & \omega_N^{-2(N-1)} I & \cdots & \omega_N^{-(N-1)(N-1)} I \end{pmatrix}, \tag{3.1}$$

I being the $r \times r$ identity matrix. Such a matrix is unitary, and the $r \times r$ block in the $(m+1)$th block row and $(n+1)$th block column of $\Omega_{r,N} C_N(A) \Omega_{r,N}^*$ equals

$$\begin{aligned} &\frac{1}{N} \sum_{l=0}^{N-1} \sum_{s=0}^{N-1} \omega_N^{-lm} \omega_N^{sn} \sum_{k \in \mathbb{Z}} A_{n-l+kN} \\ &= \frac{1}{N} \sum_{l=0}^{N-1} \omega_N^{(n-m)l} \sum_{s=0}^{N-1} \sum_{k \in \mathbb{Z}} \omega_N^{n(s-l+kN)} A_{s-l+kN} = \frac{1}{N} \left(\sum_{l=0}^{N-1} \omega_N^{(n-m)l} \right) \mathcal{A}(2\pi i n/N) \\ &= \delta_{m,n} \mathcal{A}(2\pi i n/N), \end{aligned}$$

δ being the Kronecker delta. □

Since the singular values are preserved by unitary transformations and the singular values of a block diagonal matrix are the singular values of the diagonal blocks, the theorem above has the following corollary.

COROLLARY 3.2. *A number σ is a singular value of $C_N(A)$ if and only if it is a singular value of $\mathcal{A}(2\pi i n/N)$ for some $n = 0, 1, \ldots, N-1$.*

Let us note here that the 2-norm of a matrix M equals its largest singular value, which we will denote $\sigma_{\max}(M)$. Similarly, $\sigma_{\min}(M)$ will stand for the smallest singular value, the 2-norm of M^{-1}.

COROLLARY 3.3.

$$\operatorname{cond}(C_N(A)) = \frac{\max_{n=0,\ldots,N-1} \sigma_{\max}(\mathcal{A}(2\pi i n/N))}{\min_{n=0,\ldots,N-1} \sigma_{\min}(\mathcal{A}(2\pi i n/N))}.$$

If N_1 is a divisor of N, $\operatorname{cond}_2(C_{N_1}(A)) \le \operatorname{cond}_2(C_N(A))$, because all the singular values of $C_{N_1}(A)$ are simultaneously singular values of $C_N(A)$. This means that, for the recursive transform, we could estimate the condition in each step by the condition number of the largest block circulant matrix involved, applied in the first step of the recursion, since in each next step just an m-times smaller matrix is used, m being the dilation factor.

It would be useful to have some estimate completely independent of the size of the block circulant matrix. One such estimate is straightforward,

$$\operatorname{cond}(C_N(A)) \le \frac{\sup_{\xi \in [0,2\pi)} \sigma_{\max}(\mathcal{A}(\xi))}{\inf_{\xi \in [0,2\pi)} \sigma_{\min}(\mathcal{A}(\xi))}. \tag{3.2}$$

Notice that if the curves of the largest and smallest singular values are continuous (which happens, for example, for compactly supported wavelets, when A contains only a finite number of nonzero entries), this is the lowest upper bound for $\operatorname{cond}(C_N(A))$ independent of N. We will show in the next section that for any reasonable wavelet the right-hand side of (3.2) represents, in fact, the condition number of $C(A)$.

4. Norm and condition number of block-Toeplitz operators. As in the previous section, we will apply here a "block Fourier transform." However, here the situation is a little more complicated than in the case of finite matrices.

Let us denote by $l^2_r(\mathbb{Z})$ the Hilbert space of (column) vectors of length r with all components in $l^2(\mathbb{Z})$. We can see this space also as a space of vector-valued sequences. The inner product is

$$\langle \boldsymbol{a}, \boldsymbol{b}\rangle_{l^2_r(\mathbb{Z})} = \sum_{s=1}^{r} \langle a^{(s)}, b^{(s)}\rangle_{l^2(\mathbb{Z})} = \sum_{s=1}^{r}\sum_{k\in\mathbb{Z}} a_k^{(s)}\overline{b}_k^{(s)} = \sum_{k\in\mathbb{Z}} \boldsymbol{b}_k^* \boldsymbol{a}_k;$$

subscripts determine entries of sequences, while superscripts determine entries of vectors. Similarly, $L^2_r([0,2\pi))$ is the Hilbert space of r-vectors of square integrable functions on $[0,2\pi)$ with the inner product

$$\begin{aligned}\langle \boldsymbol{f}, \boldsymbol{g}\rangle_{L^2_r([0,2\pi))} &= \sum_{s=1}^{r} \langle f^{(s)}, g^{(s)}\rangle_{L^2([0,2\pi))} = \sum_{s=1}^{r} \int_0^{2\pi} f^{(s)}(\xi)\overline{g^{(s)}(\xi)}\, d\xi \\ &= \int_0^{2\pi} \boldsymbol{g}(\xi)^* \boldsymbol{f}(\xi)\, d\xi.\end{aligned}$$

To find the norm of the operator $C(A)$ induced by the norm of $l^2(\mathbb{Z})$, we employ Hilbert space isomorphisms of these spaces. First, there is a trivial isomorphism between $l^2(\mathbb{Z})$ and $l^2_r(\mathbb{Z})$; $\{c_k\}_{k\in\mathbb{Z}} \longrightarrow \{\boldsymbol{c}_k\}_{k\in\mathbb{Z}}$, $\boldsymbol{c}_k = (\, c_{rk} \quad c_{rk+1} \quad \cdots \quad c_{rk+r-1}\,)^T$. Second, componentwise Fourier transform is a Hilbert space isomorphism $l^2_r(\mathbb{Z}) \longrightarrow L^2_r([0,2\pi))$. For a sequence $c \in l^2(\mathbb{Z})$ the Fourier transform $\widehat{c} \in L^2([0,2\pi))$ is defined as

$$\widehat{c}(\xi) = \frac{1}{\sqrt{2\pi}} \sum_{k\in\mathbb{Z}} c_k e^{-ik\xi},$$

where the sum converges in the $L^2([0,2\pi))$ sense. Since $\frac{1}{\sqrt{2\pi}} e^{-ik\xi}$, $k \in \mathbb{Z}$, is an orthonormal basis for $L^2([0,2\pi))$, the inverse mapping is given by

$$c_k = \left\langle \widehat{c}\,, \frac{1}{\sqrt{2\pi}} e^{-ik\xi} \right\rangle = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} \widehat{c}(\xi) e^{ik\xi}\, d\xi$$

and the Fourier transform as defined above is a Hilbert space isomorphism $l^2(\mathbb{Z}) \longrightarrow L^2([0,2\pi))$. The extension to the vector case is obvious.

Infinite Toeplitz matrices represent convolution operators. For sequences a, $b \in l^2(\mathbb{Z})$, the convolution $c = a * b$ has entries

$$c_l = \sum_{k\in\mathbb{Z}} a_k b_{l-k}, \qquad l \in \mathbb{Z}.$$

Convolution operators are closely related to multipliers. The link is the Fourier transform.

LEMMA 4.1. *Let a, $b \in l^2(\mathbb{Z})$ and let $a * b \in l^2(\mathbb{Z})$ or $\widehat{a}\,\widehat{b} \in L^2([0,2\pi))$. Then*

$$\widehat{a * b} = \sqrt{2\pi}\, \widehat{a}\, \widehat{b}. \tag{4.1}$$

Proof. For any $l \in \mathbb{Z}$,

$$\big(\widehat{\{\overline{b}_{l-k}\}_{k\in\mathbb{Z}}}\big)(\xi) = \sum_{k\in\mathbb{Z}} \overline{b}_{l-k} e^{-ik\xi} = \overline{e^{il\xi} \sum_{k\in\mathbb{Z}} b_k e^{-ik\xi}} = \overline{e^{il\xi}\, \widehat{b}(\xi)}.$$

Because the Fourier transform is a Hilbert space isomorphism,

$$\sum_{k\in\mathbb{Z}} a_k b_{l-k} = \langle \{a_k\}_{k\in\mathbb{Z}}, \{\overline{b}_{l-k}\}_{k\in\mathbb{Z}}\rangle_{l^2(\mathbb{Z})} = \left\langle \widehat{a},\ \overline{e^{il\xi}\widehat{b}(\xi)} \right\rangle_{L^2([0,2\pi))}$$
$$= \int_0^{2\pi} \widehat{a}(\xi)\widehat{b}(\xi)\, e^{il\xi}.$$

The last term represents the lth entry of the inverse Fourier transform of $\sqrt{2\pi}\,\widehat{a}\,\widehat{b}$. ∎

THEOREM 4.2. *The operator $l^2(\mathbb{Z}) \longrightarrow l^2(\mathbb{Z})$ represented by $C(A)$ is isomorphic with a matrix multiplier $\mathcal{A}(\xi) = \sum_{k\in\mathbb{Z}} A_k e^{ik\xi}$ that maps $L_r^2([0,2\pi)) \longrightarrow L_r^2([0,2\pi))$, $\boldsymbol{u}(\xi) \longrightarrow \mathcal{A}(\xi)\boldsymbol{u}(\xi)$.*

Proof. By the former isomorphism, $C(A)$ is isomorphic with the operator $l_r^2(\mathbb{Z}) \to l_r^2(\mathbb{Z})$, for which $\boldsymbol{d}$, the image of $\boldsymbol{c}$, is given by the formula

$$\boldsymbol{d}_l = \sum_{k\in\mathbb{Z}} A_{k-l}\boldsymbol{c}_k, \qquad l \in \mathbb{Z}.$$

We will slightly abuse the notation and denote this operator also by $C(A)$.

Since we assume that $C(A)$ represents the change from one Riesz basis to another, $\{a_n^{(k,l)}\}_{n\in\mathbb{Z}} \in l^2(\mathbb{Z})$ and the series $\mathcal{A}(\xi) = \sum_{k\in\mathbb{Z}} A_k e^{ik\xi}$ converges componentwise in the $L^2([0,2\pi))$ sense. A straightforward calculation shows that (4.1) can be extended to the matrix/vector case (the Fourier transform being defined componentwise). Because $(\widehat{\{A_{-k}\}_{k\in\mathbb{Z}}})(\xi) = \widehat{A}(-\xi)$,

$$(\widehat{C(A)\boldsymbol{c}})(\xi) = \sqrt{2\pi}\widehat{A}(-\xi)\widehat{\boldsymbol{c}}(\xi),$$

whenever $C(A)\boldsymbol{c} \in l_r^2(\mathbb{Z})$ or $\widehat{A}(-\xi)\widehat{\boldsymbol{c}}(\xi) \in L_r^2([0,2\pi))$. A convolution-type operator $C(A)$ thus becomes in the Fourier domain, indeed, the matrix multiplier $\mathcal{A}$. ∎

The norm of $C(A)$ induced by $l^2(\mathbb{Z})$ thus equals the norm of the matrix multiplier $\mathcal{A}$ as an operator $L_r^2([0,2\pi)) \longrightarrow L_r^2([0,2\pi))$. The following theorem gives formulas for the norm of a multiplier and its inverse.

THEOREM 4.3. *Let $\mathcal{A}$ be an $r \times r$ matrix multiplier with measurable components. Then*

$$\sup_{\|\boldsymbol{u}\|_{L_r^2([0,2\pi))}=1} \|\mathcal{A}\boldsymbol{u}\|_{L_r^2([0,2\pi))} = \operatorname*{ess\,sup}_{\xi\in[0,2\pi)} \max_{\|\boldsymbol{y}\|_2=1} \|\mathcal{A}(\xi)\boldsymbol{y}\|_2, \tag{4.2}$$

$$\inf_{\|\boldsymbol{u}\|_{L_r^2([0,2\pi))}=1} \|\mathcal{A}\boldsymbol{u}\|_{L_r^2([0,2\pi))} = \operatorname*{ess\,inf}_{\xi\in[0,2\pi)} \min_{\|\boldsymbol{y}\|_2=1} \|\mathcal{A}(\xi)\boldsymbol{y}\|_2. \tag{4.3}$$

Proof. Let us set

$$\Lambda = \operatorname*{ess\,sup}_{\xi\in[0,2\pi)} \max_{\|\boldsymbol{y}\|_2=1} \|\mathcal{A}(\xi)\boldsymbol{y}\|_2$$

and let $\boldsymbol{x} \in L_r^2([0,2\pi))$. Then

$$\|\mathcal{A}\boldsymbol{x}\|^2_{L_r^2([0,2\pi))} = \int_0^{2\pi} \|A(\xi)\boldsymbol{x}(\xi)\|_2^2 d\xi \le \int_0^{2\pi} \|A(\xi)\|_2^2\, \|\boldsymbol{x}(\xi)\|_2^2 d\xi \le \Lambda^2 \|\boldsymbol{x}\|^2_{L_r^2([0,2\pi))},$$

and hence

$$\sup_{\|\boldsymbol{u}\|_{L_r^2([0,2\pi))}=1} \|\mathcal{A}\boldsymbol{u}\|_{L_r^2([0,2\pi))} \le \Lambda.$$

We need to show that we have the lowest upper bound, in other words, that for each $\epsilon > 0$ we can find $\boldsymbol{x}_\epsilon$ such that $\|\boldsymbol{x}_\epsilon\|_{L^2_r([0,2\pi))} = 1$ and

$$\|\mathcal{A}\boldsymbol{x}_\epsilon\|_{L^2_r([0,2\pi))} > \Lambda - \epsilon.$$

Let us take pointwise the singular value decomposition of $\mathcal{A}$,

$$\mathcal{A}(\xi) = \mathcal{V}(\xi)^* \Sigma(\xi) \mathcal{U}(\xi),$$

where $\mathcal{V}(\xi)$ and $\mathcal{U}(\xi)$ are $r \times r$ unitary matrices and $\Sigma(\xi)$ is the diagonal matrix with the singular values of $\mathcal{A}(\xi)$ on the diagonal, in decreasing order. We will denote these singular values $\sigma_j(\xi)$, $j = 1, \ldots, r$. To construct $\boldsymbol{x}_\epsilon$ we need a path of right singular vectors corresponding to the largest singular value (something like the first column of $\mathcal{U}$), but we have to ensure that this path is square integrable.

First, $\sigma_1(\xi) = \|\mathcal{A}(\xi)\|_2$ is a measurable function, because $\mathcal{A}$ has measurable components and the matrix norm is a continuous function of the entries. Let us define

$$\begin{array}{rcll} \mathcal{C}(\xi) & = & \frac{1}{\sigma_1(\xi)^2} \mathcal{A}(\xi)^* \mathcal{A}(\xi), & \text{when } \sigma_1(\xi) \neq 0, \\ & = & I\ , & \text{otherwise.} \end{array}$$

Then $\mathcal{C}$ has measurable components and, for $k \to +\infty$, $\mathcal{C}(\xi)^k \longrightarrow \mathcal{P}(\xi)$, where

$$\mathcal{P}(\xi) = \mathcal{U}(\xi)^* \mathcal{D}(\xi) \mathcal{U}(\xi)$$

and $\mathcal{D}(\xi)$ is a diagonal matrix with the elements on the diagonal equal to either 1 or 0; if $\sigma_1(\xi)$ is of multiplicity m (m depending on ξ), then first m elements are 1 and all the others are 0. Notice that $\mathcal{P}(\xi)$ is the orthogonal projector onto the subspaces spanned by all right singular vectors corresponding to singular values $\sigma_1(\xi) = \cdots = \sigma_m(\xi)$. Because $\mathcal{P}$ is the limit of a sequence of matrices with measurable components, its components are measurable too.

Now, for any $\epsilon > 0$, the set

$$\mathcal{S}_\epsilon = \{\xi \in [0, 2\pi) :\ \sigma_1(\xi) > \Lambda - \epsilon\}$$

is a measurable set and $\mu(\mathcal{S}_\epsilon) > 0$. Since $\mathcal{P}(\xi) \neq 0$ for any ξ, there exist j and a set $\tilde{\mathcal{S}}_\epsilon \subset \mathcal{S}_\epsilon$, $\mu(\tilde{\mathcal{S}}_\epsilon) > 0$ such that $\boldsymbol{p}(\xi)$, the jth column of $\mathcal{P}(\xi)$, is nonzero for $\xi \in \tilde{\mathcal{S}}_\epsilon$. Let us set

$$\begin{array}{rcll} \boldsymbol{x}_\epsilon(\xi) & = & \frac{1}{\sqrt{\mu(\tilde{\mathcal{S}}_\epsilon)} \|\boldsymbol{p}(\xi)\|_2} \boldsymbol{p}(\xi), & \xi \in \tilde{\mathcal{S}}_\epsilon, \\ & = & 0, & \text{otherwise.} \end{array}$$

Because $\boldsymbol{x}_\epsilon$ has measurable components and $|\tilde{x}^{(s)}_\epsilon(\xi)|^2 \leq \|\boldsymbol{x}_\epsilon(\xi)\|_2^2 \in \{0, 1/\mu(\tilde{\mathcal{S}}_\epsilon)\}$, we have $\boldsymbol{x}_\epsilon \in L^2_r([0, 2\pi))$. A simple calculation shows that $\|\boldsymbol{x}_\epsilon\|_{L^2_r([0,2\pi))} = 1$. We have

$$\begin{aligned} \mathcal{A}(\xi)\mathcal{P}(\xi) &= \mathcal{V}(\xi)^* \Sigma(\xi) \mathcal{U}(\xi)\, \mathcal{U}(\xi)^* \mathcal{D}(\xi) \mathcal{U}(\xi) = \sigma_1(\xi)\, \mathcal{V}(\xi)^* \mathcal{D}(\xi) \mathcal{U}(\xi) \\ &= \sigma_1(\xi)\, \mathcal{V}(\xi)^* \mathcal{U}(\xi)\, \mathcal{P}(\xi); \end{aligned}$$

consequently,

$$\mathcal{A}(\xi)\boldsymbol{x}_\epsilon(\xi) = \sigma_1(\xi)\, \mathcal{V}(\xi)^* \mathcal{U}(\xi)\, \boldsymbol{x}_\epsilon(\xi)$$

and

$$\|\mathcal{A}(\xi)\boldsymbol{x}_\epsilon(\xi)\|_2 = \sigma_1(\xi) \|\boldsymbol{x}_\epsilon(\xi)\|_2.$$

Finally,

$$\begin{aligned}\|\mathcal{A}\boldsymbol{x}_\epsilon\|^2_{L^2_r([0,2\pi))} &= \int_0^{2\pi} \|\mathcal{A}(\xi)\boldsymbol{x}_\epsilon(\xi)\|_2^2\, d\xi = \int_0^{2\pi} \sigma_1(\xi)^2 \|\boldsymbol{x}_\epsilon(\xi)\|_2^2\, d\xi \\ &= \int_{\tilde{\mathcal{S}}_\epsilon} \frac{\sigma_1(\xi)^2}{\mu(\tilde{\mathcal{S}}_\epsilon)}\, d\xi > (\Lambda-\epsilon)^2,\end{aligned}$$

which finishes the proof of the first part of the theorem.

Now, let us concentrate on the second part of the statement. Let us denote

$$\tilde{\Lambda} = \operatorname*{ess\,inf}_{\xi\in[0,2\pi)} \min_{\|\boldsymbol{y}\|_2=1} \|\mathcal{A}(\xi)\boldsymbol{y}\|_2.$$

Clearly, for any $\boldsymbol{u}$, $\|\boldsymbol{u}\|_{L^2_r([0,2\pi))} = 1$,

$$\begin{aligned}\|\mathcal{A}\boldsymbol{u}\|_{L^2_r([0,2\pi))} &= \int_0^{2\pi} \|A(\xi)\boldsymbol{u}(\xi)\|_2^2 d\xi \geq \int_0^{2\pi} \left(\min_{\|\boldsymbol{y}\|_2=1} \|A(\xi)\boldsymbol{y}\|_2\right)^2 \|\boldsymbol{u}(\xi)\|_2^2 d\xi \\ &\geq \tilde{\Lambda}^2 \|\boldsymbol{u}\|^2_{L^2_r([0,2\pi))} = \tilde{\Lambda}^2.\end{aligned}$$

We now have to show that for every $\epsilon > 0$ there exists $\boldsymbol{x}_\epsilon$, $\|\boldsymbol{x}_\epsilon\|_{L^2_r([0,2\pi))} = 1$, such that

$$\|\mathcal{A}\boldsymbol{x}_\epsilon\|_{L^2_r([0,2\pi))} < \tilde{\Lambda} + \epsilon.$$

In order to do that we first need to construct a square integrable path of right singular vectors corresponding to the path of the smallest singular values, σ_r.

Let us take, again, the pointwise singular value decomposition of $\mathcal{A}$,

$$\mathcal{A}(\xi) = \mathcal{V}^*(\xi)\Sigma(\xi)\mathcal{U}(\xi).$$

Now, for a positive integer k, let us consider a matrix $\mathcal{A}(\xi)^*\mathcal{A}(\xi) + \frac{1}{k}I$. We have

$$\mathcal{A}(\xi)^*\mathcal{A}(\xi) + \frac{1}{k}I = \mathcal{U}(\xi)^* \left(\Sigma(\xi)^2 + \frac{1}{k}I\right)\mathcal{U}(\xi);$$

therefore, such a matrix is invertible and the norm of the inverse is $(\sigma_r(\xi)^2 + \frac{1}{k})^{-1}$. If we set

$$\mathcal{C}_k(\xi) = \left(\sigma_r(\xi)^2 + \frac{1}{k}\right)\left(\mathcal{A}(\xi)^*\mathcal{A}(\xi) + \frac{1}{k}I\right)^{-1},$$

then $\mathcal{C}_k$ has measurable components and $\mathcal{C}_k(\xi)^l \longrightarrow \mathcal{P}(\xi)$, $k \to +\infty$, $l \to +\infty$, where

$$\mathcal{P}(\xi) = \mathcal{U}(\xi)^*\mathcal{D}(\xi)\mathcal{U}(\xi);$$

$\mathcal{D}(\xi)$ is, again, a diagonal matrix with the elements on the diagonal equal to either 1 or 0. But now, if $\sigma_r(\xi)$ is of multiplicity m, then the first $r-m$ elements are 0 and all the others are 1. The components of the matrix $\mathcal{P}$ are measurable functions and the matrix $\mathcal{P}(\xi)$ is now the orthogonal projector onto the subspace spanned by right singular vectors corresponding to the singular values $\sigma_{r-m+1}(\xi) = \cdots = \sigma_r(\xi)$ and, for any vector $\boldsymbol{x}(\xi)$ of unit norm from its range,

$$\|\mathcal{A}(\xi)\boldsymbol{x}\|_2 = \sigma_r(\xi) = \min_{\|\boldsymbol{y}\|_2=1} \|\mathcal{A}(\xi)\boldsymbol{y}\|_2.$$

The rest of the proof would follow the lines of the proof of the first part, with $\mathcal{S}_\epsilon$ being chosen as

$$\mathcal{S}_\epsilon = \{\xi \in [0, 2\pi) : \ \sigma_r(\xi) < \tilde{\Lambda} + \epsilon\};$$

we would have

$$\begin{aligned}\|\mathcal{A}\boldsymbol{x}_\epsilon\|^2_{L^2_r([0,2\pi))} &= \int_0^{2\pi} \|\mathcal{A}(\xi)\boldsymbol{x}_\epsilon(\xi)\|_2^2 \, d\xi = \int_0^{2\pi} \sigma_r(\xi)^2 \|\boldsymbol{x}_\epsilon(\xi)\|_2^2 \, d\xi \\ &= \int_{\tilde{\mathcal{S}}_\epsilon} \frac{\sigma_r(\xi)^2}{\mu(\tilde{\mathcal{S}}_\epsilon)} \, d\xi < (\tilde{\Lambda} + \epsilon)^2. \qquad \square\end{aligned}$$

The norm of $\mathcal{A}$ induced by the norm of $L^2_r([0, 2\pi))$ thus is

$$\|\mathcal{A}\|_{\mathcal{B}(L^2_r([0,2\pi)))} = \operatorname*{ess\,sup}_{\xi\in[0,2\pi)} \ \sigma_{\max}(\mathcal{A}(\xi));$$

the mapping is invertible if and only if

$$0 < \operatorname*{ess\,inf}_{\xi\in[0,2\pi)} \ \min_{\|\boldsymbol{y}\|_2=1} \|\mathcal{A}(\xi)\boldsymbol{y}\|_2$$

and the norm of the inverse equals

$$\|\mathcal{A}^{-1}\|_{\mathcal{B}(L^2_r([0,2\pi)))} = \operatorname*{ess\,inf}_{\xi\in[0,2\pi)} \sigma_{\min}(\mathcal{A}(\xi)).$$

Combining the results above we obtain the following theorem.

THEOREM 4.4. *The condition of the operator $C(A)$ (in the norm induced by the norm of $l^2(\mathbb{Z})$) is*

$$\operatorname{cond}(C(A)) = \frac{\operatorname{ess\,sup}_{\xi\in[0,2\pi)} \sigma_{\max}(\mathcal{A}(\xi))}{\operatorname{ess\,inf}_{\xi\in[0,2\pi)} \sigma_{\min}(\mathcal{A}(\xi))}.$$

For all wavelets of practical interest, A has only a finite number of nonzero entries or at least the sequences forming its rows decay very fast. This implies some smoothness of entries of $\mathcal{A}$ and, consequently, the essential supremum of $\sigma_{\max}$ and essential infimum of $\sigma_{\min}$ coincide with the supremum and infimum, respectively. As we already pointed out, $\operatorname{cond}(C(A))$ then represents $\sup_N \operatorname{cond}(C_N(A))$.

Let us make a few comments about the structure of singular values of $\mathcal{A}(\xi)$ in relation to some special properties of A. First, when the underlying bases comprise real functions, the entries of A are real and, consequently, $\mathcal{A}(\xi) = \overline{\mathcal{A}(2\pi - \xi)}$. This means that the singular values in $\pi - \xi$ and $\pi + \xi$ are the same, and we can restrict our attention onto interval $[0, \pi]$, only.

Another interesting effect is caused by all the scaling and wavelet functions and their biorthogonal counterparts being compactly supported. This corresponds to the fact that only a finite number of square blocks both in A and in $\tilde{A}$ that generate $C(\tilde{A}) = (C(A)^*)^{-1}$ are nonzero. It is well known, particularly in the filter bank context (see, e.g., [12], [13]), that this happens if and only if there exist a nonzero constant α and an integer p such that

$$\det\left(\sum_{k\in\mathbb{Z}} A_k z^{-k}\right) = \alpha z^{-p}$$

for any $z \in \mathbb{C}$, $z \neq 0$. Because the determinant is the product of singular values, the equation above implies that

$$\prod_j \sigma_j(\mathcal{A}(\xi)) = \beta$$

for some positive constant β independent of ξ. This is particularly useful when A has only two rows. The singular values of $\mathcal{A}(\xi)$ are then inversely proportional and the maximum and minimum over ξ then occur at the same point. That is,

$$\operatorname{cond}(C(A)) = \max_{\xi\in[0,2\pi)} \operatorname{cond}(\mathcal{A}(\xi)).$$

5. Alternative expression. Let the sequences that form the rows of A be $\{b_n^{(s)}\}_{n\in\mathbb{Z}}$, $s = 1, \dots, r$; $b_0^{(s)} = a_0^{(s,1)}$. Sometimes it is easier to deal with Fourier series

$$b^{(s)}(\xi) = \frac{1}{\sqrt{r}} \sum_{n\in\mathbb{Z}} b_n^{(s)} e^{-in\xi}, \qquad s = 1, \dots, r$$

than with $\mathcal{A}$. We will see an example in section 6, when we will study conditioning of biorthogonal spline wavelets. In these cases it is better to use a different matrix function.

THEOREM 5.1. *A number σ is a singular value of $\mathcal{A}(\xi)$ if and only if it is a singular value of $\mathcal{B}(-\xi/r)$, where*

$$\mathcal{B}(\xi) = \begin{pmatrix} b^{(1)}(\xi) & b^{(1)}(\xi + \frac{2\pi}{r}) & \cdots & b^{(1)}(\xi + \frac{(r-1)2\pi}{r}) \\ b^{(2)}(\xi) & b^{(2)}(\xi + \frac{2\pi}{r}) & \cdots & b^{(2)}(\xi + \frac{(r-1)2\pi}{r}) \\ \vdots & \vdots & \cdots & \vdots \\ b^{(r)}(\xi) & b^{(r)}(\xi + \frac{2\pi}{r}) & \cdots & b^{(r)}(\xi + \frac{(r-1)2\pi}{r}) \end{pmatrix}.$$

Proof. Using the notation introduced in the proof of Theorem 3.1, we have, for any $s = 1, \dots, r$ and any $k = 0, \dots, r-1$,

$$\begin{aligned} \frac{1}{\sqrt{r}} \sum_{l=0}^{r-1} b^{(s)}\left(\xi + \frac{2\pi l}{r}\right) \omega_r^{lk} &= \frac{1}{r} \sum_{l=0}^{r-1} \sum_{n\in\mathbb{Z}} b_n^{(s)} e^{-in\xi} e^{-il(n-k)2\pi/r} \\ &= \frac{1}{r} \sum_{n\in\mathbb{Z}} b_n^{(s)} e^{-in\xi} \left(\sum_{l=0}^{r-1} \omega_r^{l(n-k)} \right) \\ &= \sum_{n\in\mathbb{Z}} b_{nr+k}^{(s)} e^{-i(nr+k)\xi} = e^{-ik\xi} \sum_{n\in\mathbb{Z}} a_n^{(s,k+1)} e^{-in(r\xi)}. \end{aligned}$$

This is because $\sum_{l=0}^{r-1} \omega_r^{l(n-k)}$ equals r if $n-k$ is divisible by r, and it is 0 otherwise. Consequently,

$$\mathcal{B}(\xi)\Omega_{1,r}^* = \sqrt{r}\mathcal{A}(-r\xi)D_r(\xi),$$

where $\Omega_{1,r}$ is the $r \times r$ matrix of the discrete Fourier transform and $D_r(\xi)$ is the diagonal matrix with the diagonal entries equal to $e^{-ik\xi}$, $k = 0, \dots, r-1$ (in this particular order). Since $\Omega_{1,r}$ is unitary and so is $D_r(\xi)$ (for any ξ), the statement of the theorem holds. ∎

Just let us point out here that, instead of considering each row of A separately, we could use block rows, each of them comprising, say, p rows. We would then obtain similar results with some $p \times p$ matrices $B_n^{(s)}$ instead of scalars $b_n^{(s)}$; instead of $\Omega_{1,r}$ we would use $\Omega_{p,r/p}$ and, similarly, $D_r(\xi)$ would be replaced by a matrix with $p \times p$ diagonal blocks equal to $e^{-ik\xi} I$, $k = 0, \ldots, r/p$. This might be useful for the case of multiwavelets (more than one scaling function) when the two-scale equations analogous to (2.1) have matrix coefficients (see, e.g., [10]).

6. Conditioning of biorthogonal wavelets based on B-splines. Biorthogonal wavelets based on B-splines were introduced by Cohen, Daubechies, and Feauveau in [5]. To the B-spline basis function of a particular order (which represents a scaling function) there exists a whole family of possible biorthogonal counterparts with different size of support and regularity. We will use here the notation of [2], where the sequences determining the scaling and wavelet functions through the two-scale equations of type (2.1) are given in terms of trigonometric polynomials:

$$m_0(\xi) = \sqrt{2} \sum_k h_k e^{-ik\xi}, \qquad \tilde{m}_0(\xi) = \sqrt{2} \sum_k \tilde{h}_k e^{-ik\xi},$$
$$m_1(\xi) = \sqrt{2} \sum_k g_k e^{-ik\xi}, \qquad \tilde{m}_1(\xi) = \sqrt{2} \sum_k \tilde{g}_k e^{-ik\xi}.$$

Since the scaling function φ equals the B-spline of order n,

$$m_0(\xi) = \left(\frac{1 + e^{-i\xi}}{2} \right)^{n+1}.$$

For any integer K such that $2K \geq n + 1$,

$$\tilde{m}_0(\xi) = \cos^{2K}(\xi/2) P_K(\sin^2(\xi/2)) \left(\frac{1 + e^{i\xi}}{2} \right)^{-n-1}$$

determines a biorthogonal scaling function; P_K is the solution of the Bezout problem

$$y^K P_K(1 - y) + (1 - y)^K P_K(y) = 1; \tag{6.1}$$

in particular,

$$P_K(y) = \sum_{j=0}^{K-1} \binom{K - 1 + j}{j} y^j.$$

The trigonometric polynomials m_1 and $\tilde{m}_1$ corresponding to the wavelet filters are then defined as

$$m_1(\xi) = e^{-i\xi} \, \overline{\tilde{m}_0(\xi + \pi)}, \qquad \tilde{m}_1(\xi) = e^{-i\xi} \, \overline{m_0(\xi + \pi)}. \tag{6.2}$$

We have

$$\mathcal{B}(\xi) = \begin{pmatrix} m_0(\xi) & m_0(\xi + \pi) \\ m_1(\xi) & m_1(\xi + \pi) \end{pmatrix}$$

and, because we deal with compactly supported real classical wavelets with dilations by 2, we are interested in the maximum of the condition number of $\mathcal{B}(\xi)$ on $[0, \pi/2]$.

TABLE 1
Condition numbers for spline biorthogonal wavelets of order n; $K = \lceil \frac{n+1}{2} \rceil + L$.

	n	0	1	2	3	4	5	6	7	8	9	10 $\cdot 10^3$	11 $\cdot 10^3$	12 $\cdot 10^3$
L														
0		1.00	2.62	4.00	10.9	16.0	102.	91.6	1227.	834.	15878.	8.77	213.	99.4
1		1.28	2.00	4.00	8.00	16.0	32.0	64.0	166.	333.	1336.	2.51	12.5	22.0
2		1.42	2.00	4.00	8.00	16.0	32.0	64.0	128.	259.	691.	1.47	5.06	10.6
3		1.52	2.00	4.00	8.00	16.0	32.0	64.0	128.	256.	536.	1.13	3.16	6.89
4		1.59	2.00	4.00	8.00	16.0	32.0	64.0	128.	256.	512.	1.03	2.43	5.29
5		1.64	2.00	4.00	8.00	16.0	32.0	64.0	128.	256.	512.	1.02	2.13	4.53
6		1.68	2.00	4.00	8.00	16.0	32.0	64.0	128.	256.	512.	1.02	2.05	4.18
7		1.72	2.03	4.00	8.00	16.0	32.0	64.0	128.	256.	512.	1.02	2.05	4.10
8		1.75	2.07	4.00	8.00	16.0	32.0	64.0	128.	256.	512.	1.02	2.05	4.10
9		1.78	2.11	4.00	8.00	16.0	32.0	64.0	128.	256.	512.	1.02	2.05	4.10
10		1.80	2.15	4.00	8.00	16.0	32.0	64.0	128.	256.	512.	1.02	2.05	4.10
11		1.82	2.19	4.00	8.00	16.0	32.0	64.0	128.	256.	512.	1.02	2.05	4.10
12		1.84	2.23	4.00	8.00	16.0	32.0	64.0	128.	256.	512.	1.02	2.05	4.10
13		1.86	2.26	4.00	8.00	16.0	32.0	64.0	128.	256.	512.	1.02	2.05	4.10
14		1.88	2.29	4.00	8.00	16.0	32.0	64.0	128.	256.	512.	1.02	2.05	4.10
15		1.89	2.32	4.00	8.00	16.0	32.0	64.0	128.	256.	512.	1.02	2.05	4.10
16		1.91	2.35	4.00	8.00	16.0	32.0	64.0	128.	256.	512.	1.02	2.05	4.10
17		1.92	2.38	4.01	8.00	16.0	32.0	64.0	128.	256.	512.	1.02	2.05	4.10
18		1.93	2.41	4.03	8.00	16.0	32.0	64.0	128.	256.	512.	1.02	2.05	4.10
19		1.95	2.43	4.05	8.00	16.0	32.0	64.0	128.	256.	512.	1.02	2.05	4.10
20		1.96	2.45	4.07	8.00	16.0	32.0	64.0	128.	256.	512.	1.02	2.05	4.10

TABLE 2
Condition numbers for spline biorthogonal wavelets of order n, $K = \lceil \frac{n+1}{2} \rceil + L$, optimally scaled wavelet.

	n	0	1	2	3	4	5	6	7	8	9	10	11	12
L														
0		1.00	2.41	2.00	6.16	4.39	20.1	12.2	70.0	38.1	252.	125.	924.	425.
1		1.28	1.41	2.00	2.83	4.00	5.66	8.00	13.7	18.7	39.2	51.7	121.	153.
2		1.42	1.41	2.00	2.83	4.00	5.66	8.00	11.3	16.1	26.6	38.6	72.2	104.
3		1.50	1.41	2.00	2.83	4.00	5.66	8.00	11.3	16.0	23.2	33.8	56.5	83.3
4		1.57	1.41	2.00	2.83	4.00	5.66	8.00	11.3	16.0	22.6	32.1	49.4	72.9
5		1.62	1.41	2.00	2.83	4.00	5.66	8.00	11.3	16.0	22.6	32.0	46.3	67.4
6		1.66	1.43	2.00	2.83	4.00	5.66	8.00	11.3	16.0	22.6	32.0	45.3	64.7
7		1.69	1.46	2.00	2.83	4.00	5.66	8.00	11.3	16.0	22.6	32.0	45.3	64.0
8		1.72	1.48	2.00	2.83	4.00	5.66	8.00	11.3	16.0	22.6	32.0	45.3	64.0
9		1.74	1.51	2.00	2.83	4.00	5.66	8.00	11.3	16.0	22.6	32.0	45.3	64.0
10		1.76	1.53	2.00	2.83	4.00	5.66	8.00	11.3	16.0	22.6	32.0	45.3	64.0
11		1.78	1.55	2.00	2.83	4.00	5.66	8.00	11.3	16.0	22.6	32.0	45.3	64.0
12		1.80	1.57	2.00	2.83	4.00	5.66	8.00	11.3	16.0	22.6	32.0	45.3	64.0
13		1.81	1.59	2.00	2.83	4.00	5.66	8.00	11.3	16.0	22.6	32.0	45.3	64.0
14		1.83	1.60	2.00	2.83	4.00	5.66	8.00	11.3	16.0	22.6	32.0	45.3	64.0
15		1.84	1.62	2.00	2.83	4.00	5.66	8.00	11.3	16.0	22.6	32.0	45.3	64.0
16		1.85	1.63	2.01	2.83	4.00	5.66	8.00	11.3	16.0	22.6	32.0	45.3	64.0
17		1.87	1.65	2.01	2.83	4.00	5.66	8.00	11.3	16.0	22.6	32.0	45.3	64.0
18		1.88	1.66	2.02	2.83	4.00	5.66	8.00	11.3	16.0	22.6	32.0	45.3	64.0
19		1.89	1.67	2.03	2.83	4.00	5.66	8.00	11.3	16.0	22.6	32.0	45.3	64.0
20		1.89	1.68	2.04	2.83	4.00	5.66	8.00	11.3	16.0	22.6	32.0	45.3	64.0

TABLE 3
Optimal scaling parameters.

n	0 /10	1 /10	2 /10	3 /10	4 /10	5 $/10^2$	6 $/10^2$	7 $/10^2$	8 $/10^2$	9 $/10^3$	10 $/10^3$	11 $/10^3$	12 $/10^3$
L													
0	10.0	7.07	5.00	3.16	2.40	9.95	9.16	2.86	2.99	7.93	9.20	2.18	2.72
1	9.70	7.07	5.00	3.54	2.50	17.7	12.5	7.33	5.35	25.5	19.4	8.30	6.56
2	9.42	7.07	5.00	3.54	2.50	17.7	12.5	8.83	6.20	37.6	25.9	13.9	9.65
3	9.22	7.07	5.00	3.54	2.50	17.7	12.5	8.84	6.25	43.1	29.6	17.7	12.0
4	9.07	7.07	5.00	3.54	2.50	17.7	12.5	8.84	6.25	44.2	31.2	20.2	13.7
5	8.95	7.07	5.00	3.54	2.50	17.7	12.5	8.84	6.25	44.2	31.3	21.7	14.9
6	8.85	6.99	5.00	3.54	2.50	17.7	12.5	8.84	6.25	44.2	31.3	22.1	15.5
7	8.77	6.86	5.00	3.54	2.50	17.7	12.5	8.84	6.25	44.2	31.3	22.1	15.6
8	8.70	6.74	5.00	3.54	2.50	17.7	12.5	8.84	6.25	44.2	31.3	22.1	15.6
9	8.64	6.64	5.00	3.54	2.50	17.7	12.5	8.84	6.25	44.2	31.3	22.1	15.6
10	8.58	6.54	5.00	3.54	2.50	17.7	12.5	8.84	6.25	44.2	31.3	22.1	15.6
11	8.54	6.45	5.00	3.54	2.50	17.7	12.5	8.84	6.25	44.2	31.3	22.1	15.6
12	8.49	6.37	5.00	3.54	2.50	17.7	12.5	8.84	6.25	44.2	31.3	22.1	15.6
13	8.45	6.30	5.00	3.54	2.50	17.7	12.5	8.84	6.25	44.2	31.3	22.1	15.6
14	8.42	6.24	5.00	3.54	2.50	17.7	12.5	8.84	6.25	44.2	31.3	22.1	15.6
15	8.39	6.18	5.00	3.54	2.50	17.7	12.5	8.84	6.25	44.2	31.3	22.1	15.6
16	8.36	6.13	4.98	3.54	2.50	17.7	12.5	8.84	6.25	44.2	31.3	22.1	15.6
17	8.33	6.08	4.97	3.54	2.50	17.7	12.5	8.84	6.25	44.2	31.3	22.1	15.6
18	8.30	6.03	4.95	3.54	2.50	17.7	12.5	8.84	6.25	44.2	31.3	22.1	15.6
19	8.28	5.99	4.93	3.54	2.50	17.7	12.5	8.84	6.25	44.2	31.3	22.1	15.6
20	8.26	5.95	4.91	3.54	2.50	17.7	12.5	8.84	6.25	44.2	31.3	22.1	15.6

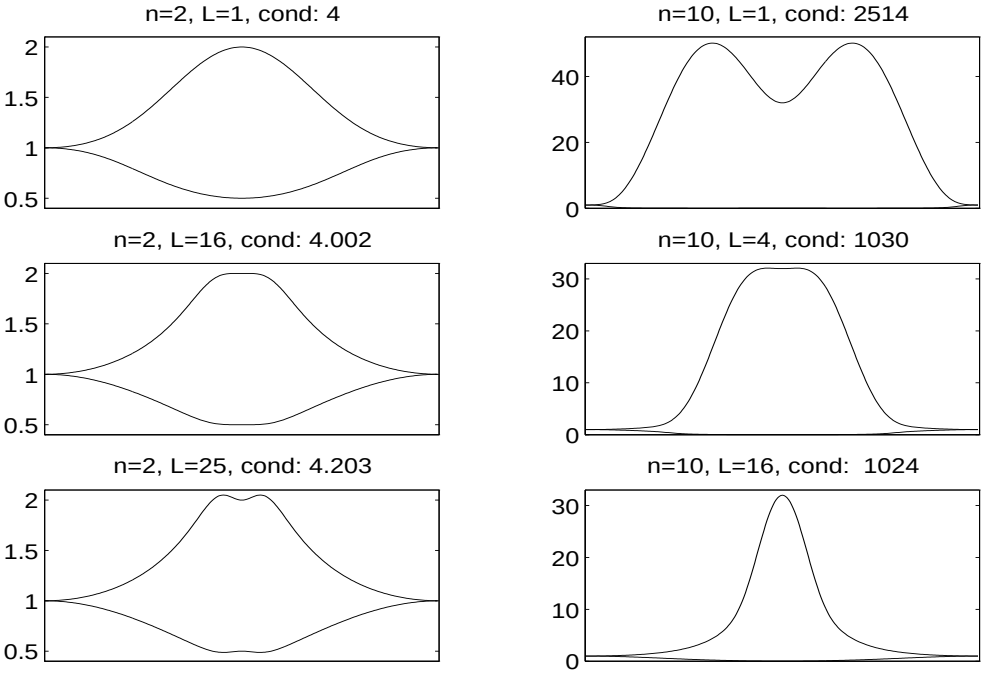


FIG. 1. *Typical shapes of the curves of singular values of* $\mathcal{A}(\xi)$, $\xi \in [0, 2\pi)$, *for spline biorthogonal wavelets (order* n, $K = \lceil \frac{n+1}{2} \rceil + L$, *unscaled).*

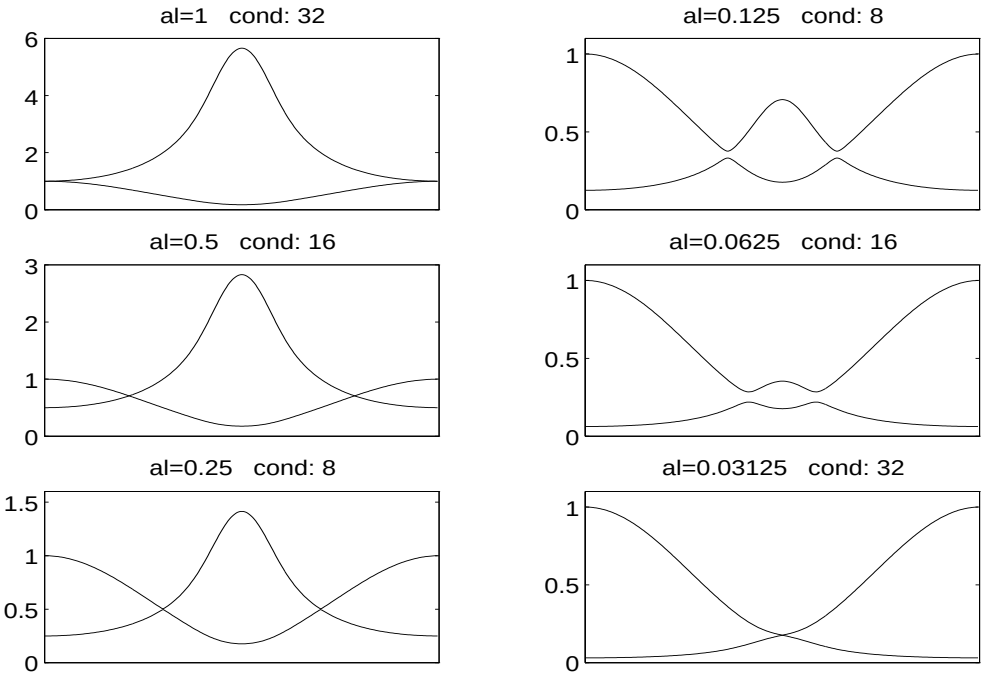


FIG. 2. *Singular value curves for different scaling parameters α; order of spline $n = 5$, $K = 13$.*

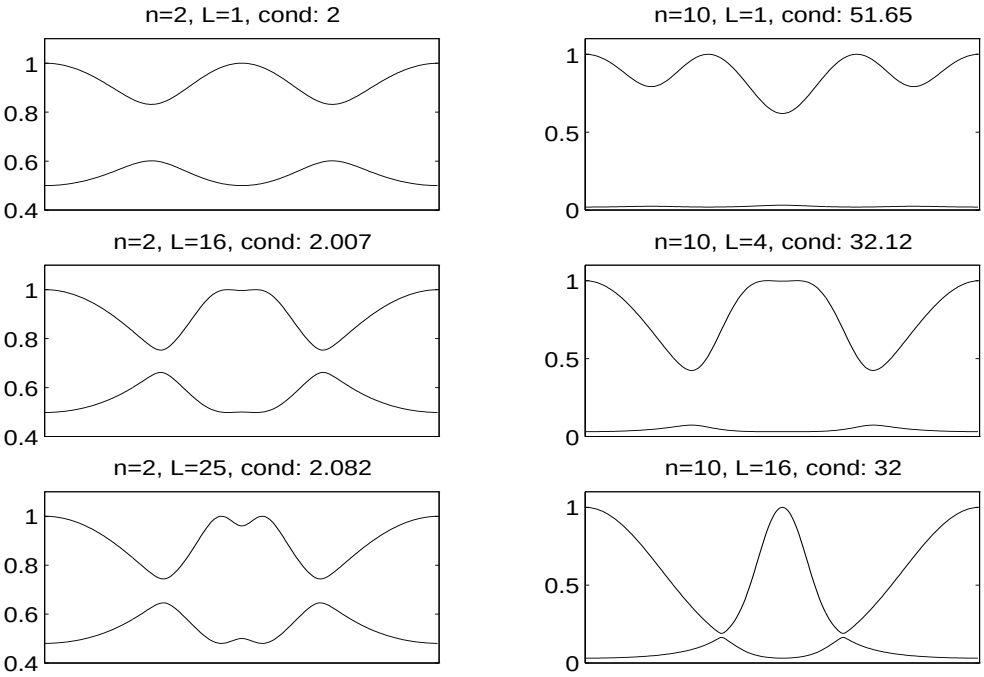


FIG. 3. *Singular value curves for optimally scaled spline biorthogonal wavelets.*

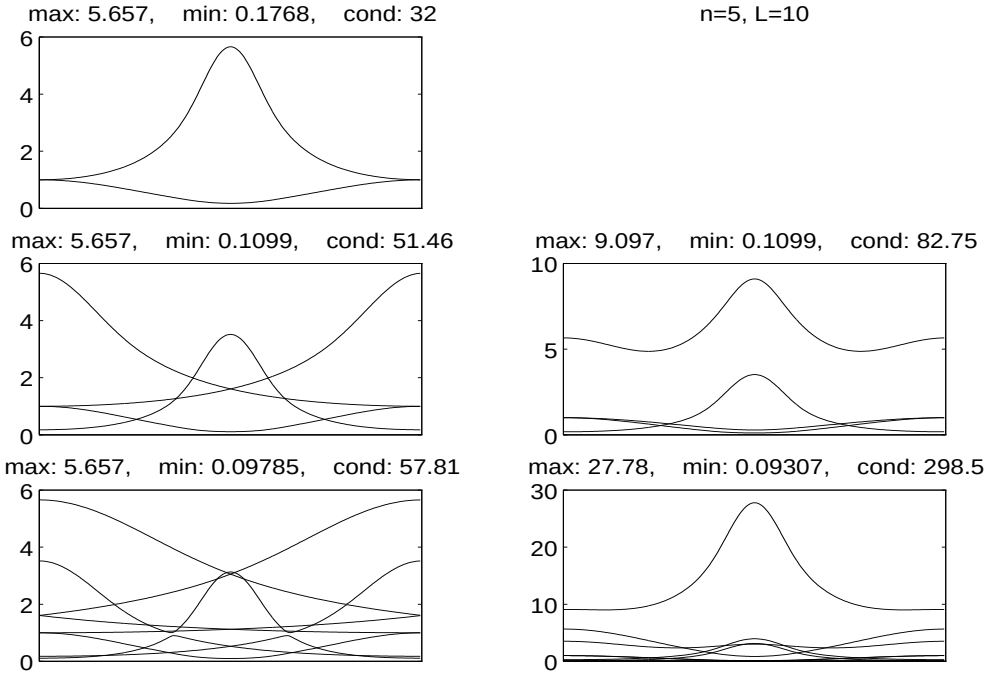


FIG. 4. *Singular value curves for classical wavelet (left) and full tree wavelet packet transform (right) of depth* $1, 2, 3$ *(from top to bottom); order of spline* $n = 5$*,* $K = 13$*, unscaled.*

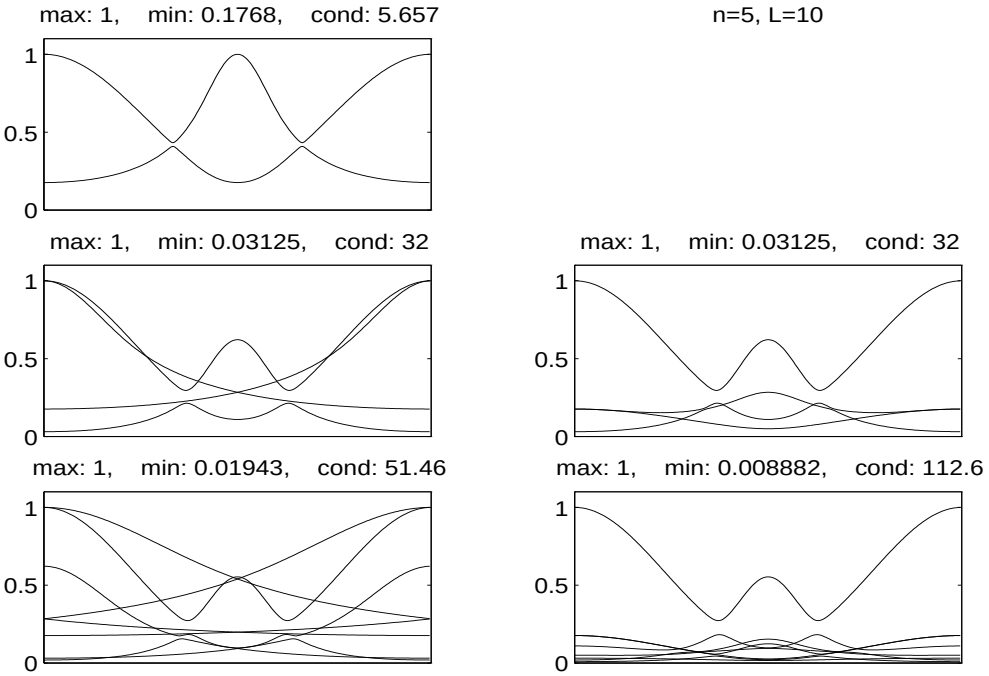


FIG. 5. *Singular value curves for classical wavelet (left) and full tree wavelet packet transform (right) of depth* $1, 2, 3$ *(from top to bottom); order of spline* $n = 5$*,* $K = 13$*, optimal scaling for depth* 1.

Squares of the singular values of $\mathcal{B}(\xi)$ are the eigenvalues of the matrix $\mathcal{B}(\xi)\mathcal{B}(\xi)^*$, and they satisfy a quadratic equation

$$\lambda^2 - \operatorname{tr}(\mathcal{B}(\xi)\mathcal{B}(\xi)^*)\lambda + \det(\mathcal{B}(\xi)\mathcal{B}(\xi)^*) = 0.$$

Fairly straightforward, although somewhat tedious, calculations show that the coefficients of this equation are

$$\det(\mathcal{B}(\xi)\mathcal{B}(\xi)^*) = |\det(\mathcal{B}(\xi))|^2 = |-e^{-i\xi}|^2 = 1$$

and

$$\operatorname{tr}(\mathcal{B}(\xi)\mathcal{B}(\xi)^*) = |m_0(\xi)|^2 + |m_0(\xi+\pi)|^2 + |m_1(\xi)|^2 + |m_1(\xi+\pi)|^2,$$

where

$$\begin{aligned} |m_0(\xi)|^2 &= (\cos^2(\xi/2))^{n+1}, \quad |m_1(\xi)|^2 = (\sin^2(\xi/2))^{2K-n-1} P_K^2(\cos^2(\xi/2)), \\ |m_0(\xi+\pi)|^2 &= (\sin^2(\xi/2))^{n+1}, \quad |m_1(\xi+\pi)|^2 = (\cos^2(\xi/2))^{2K-n-1} P_K^2(\sin^2(\xi/2)). \end{aligned}$$

THEOREM 6.1. *The numerical condition of one level of the (fast) discrete wavelet transform based on B-spline biorthogonal wavelets of order n defined above is at least 2^n, independently of the value of K.*

Proof. Since

$$P_K(\cos^2(\pi/4)) = P_K(\sin^2(\pi/4)) = P_K(1/2) = 2^{K-1}$$

(cf. (6.1)), substituting $\xi = \pi/2$ into the formulae above we obtain

$$\operatorname{tr}(\mathcal{B}(\pi/2)\mathcal{B}(\pi/2)^*) = 2^{-n} + 2^n.$$

Squares of the singular values of $\mathcal{B}(\pi/2)$ thus equal 2^n and 2^{-n}, respectively, and the condition of this matrix is 2^n. The condition hence must be at least 2^n. ∎

Numerical experiments show that the condition number often *equals* 2^n. From the point of view of conditioning, it is better to choose K smaller for low order splines and larger for higher order splines; see Table 1.

Once the scaling filters m_0 and $\tilde{m}_0$ are given, (6.2) is not the only possibility for the corresponding wavelet filters. The entire freedom can be described as follows:

$$m_1(\xi) = \alpha e^{-i(2k+1)\xi}\,\overline{\tilde{m}_0(\xi+\pi)}, \qquad \tilde{m}_1(\xi) = (1/\alpha) e^{-i(2k+1)\xi}\,\overline{m_0(\xi+\pi)},$$

$k \in \mathbb{Z}$, $\alpha \neq 0$. The choice of k is, from the point of view of the numerical condition, irrelevant, but the scaling by α can be used to improve the condition. In the case of the spline wavelets improvement can be significant. However, it turns out that whatever scaling we choose, we can't beat the exponential growth with the order of the spline.

THEOREM 6.2. *For any scaling factor α, the condition of one step of a discrete wavelet transform with a spline biorthogonal wavelet of order n is at least $2^{n/2}$.*

Proof. Instead of the condition of $\mathcal{B}(\xi)$ we need to study here the condition of

$$\mathcal{B}_\alpha(\xi) = \begin{pmatrix} m_0(\xi) & m_0(\xi+\pi) \\ \alpha m_1(\xi) & \alpha m_1(\xi+\pi) \end{pmatrix},$$

where $m_1(\xi) = e^{-i\xi}\,\overline{\tilde{m}_0(\xi+\pi)}$ as before, in (6.2). For $\xi = \pi/2$,

$$\operatorname{tr}(\mathcal{B}_\alpha(\pi/2)\mathcal{B}_\alpha(\pi/2)^*) = 2^{-n} + |\alpha|^2\, 2^n,$$

the singular values of $\mathcal{B}_\alpha(\pi/2)$ are $|\alpha|\,2^{n/2}$ and $2^{-n/2}$ and its condition hence is $|\alpha|\,2^n$ for $|\alpha| \geq 2^{-n}$ and $1/(|\alpha|2^n)$ for $|\alpha| < 2^{-n}$. On the other hand, for $\xi = 0$,

$$\mathcal{B}_\alpha(0) = \begin{pmatrix} 1 & 0 \\ 0 & -\alpha \end{pmatrix},$$

and its condition is $|\alpha|$ for $|\alpha| \geq 1$ and $1/|\alpha|$ for $|\alpha| < 1$. Combining these results we see that the condition of the wavelet transform cannot be better than $|\alpha|2^n$ if $|\alpha| \geq 2^{-n/2}$, and $1/|\alpha|$ if $|\alpha| < 2^{-n/2}$. Consequently, whatever $|\alpha|$ we choose, the condition is at least $2^{n/2}$. □

The optimal scaling parameter is usually equal or close to $2^{-n/2}$; see Tables 2 and 3. Notice that this is true especially for the wavelets that have condition number equal to 2^n. The condition of the optimally scaled wavelet then equals $2^{n/2}$, in most cases.

Figures 1–5 show some typical behavior of the singular value curves depending on the order of the spline, parameter K, scaling of the wavelet, and depth of the transform. There are some interesting details here, for example, the presence of points where the plot looks almost as if two curves were intersecting each other, while, in fact, we have two different curves that have turning points and are well separated.

Acknowledgments. When working on this paper the author was a postgraduate research scholar supported by the Australian Government. She thanks her advisor, Jaroslav Kautsky, for suggesting the topic and for many fruitful discussions.

REFERENCES

[1] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, *Image coding using wavelet transform*, IEEE Trans. Image Process., 1 (1992), pp. 205–220.

[2] A. Cohen, *Biorthogonal wavelets*, in Wavelets: A Tutorial in Theory and Applications, C. Chui, ed., Academic Press, New York, 1992, pp. 123–152.

[3] A. Cohen and I. Daubechies, *A stability criterion for biorthogonal wavelet bases and their related subband coding scheme*, Duke Math. J., 68 (1992), pp. 313–335.

[4] A. Cohen and I. Daubechies, *On the instability of arbitrary biorthogonal wavelet packets*, SIAM J. Math. Anal., 24 (1993), pp. 1340–1354.

[5] A. Cohen, I. Daubechies, and J. C. Feauveau, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 45 (1992), pp. 485–560.

[6] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[7] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA, 1991.

[8] F. Keinert, *Numerical stability of biorthogonal wavelet transforms*, Adv. Comput. Math., 4 (1995), pp. 1–26 (special issue on multiscale techniques).

[9] G. Strang, *Inner Products and Condition Numbers for Wavelets and Filter Banks*, 1994, manuscript.

[10] G. Strang and V. Strela, *Short wavelets and matrix dilation equations*, IEEE Trans. Signal Process., 43 (1995), pp. 108–115.

[11] A. Uhl, *Compact image coding using wavelets and wavelet packets based on non-stationary and inhomogeneous multiresolution analyses*, in Mathematical Imaging: Wavelet Applications in Signal and Image Processing II, Proc. SPIE 2303, A. F. Laine and M. Unser, eds., 1994, pp. 378–388.

[12] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Signal Processing Series, Prentice–Hall, Englewood Cliffs, NJ, 1992.

[13] M. Vetterli and D. Le Gall, *Perfect reconstruction FIR filter banks: Some properties and factorizations*, IEEE Trans. Acoust., Speech Signal Process., 37 (1989), pp. 1057–1071.

[14] M. V. Wickerhauser, *High-Resolution Still Picture Compression*, Tech. report, Dept. of Mathematics, Washington University, St. Louis, MO, 1992.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 4, pp. 1000–1012, October 1997

013

ON SCALED ALMOST-DIAGONAL HERMITIAN MATRIX PAIRS*

VJERAN HARI† AND ZLATKO DRMAČ‡

Abstract. This paper contains estimates concerning the block structure of Hermitian matrices H and M, which make a scaled diagonally dominant definite pair. The obtained bounds are expressed in terms of relative gaps in the spectrum of the pair (H, M) and norms of certain blocks of the matrices DHD and DMD, where D is either $[|\mathrm{diag(H)}|]^{-1/2}$ or $[\mathrm{diag(M)}]^{-1/2}$. If either of the matrices H, M is diagonal, the new results assume simple and applicable form. For scaled diagonally dominant Hermitian matrices, the new estimates compare favorably with the existing ones for accurate location of the smallest eigenvalues.

Key words. almost-diagonal matrices, scaled matrices, eigenvalue location

AMS subject classifications. 65F15, 65G05

PII. S0895479894278472

Introduction. An almost-diagonal Hermitian matrix has several important properties. First, its diagonal elements approximate its eigenvalues with an error which is quadratic with respect to the average off-diagonal element. Second, in the case of multiple eigenvalues, the matrix has a special block structure: those off-diagonal elements which link the diagonals that approximate the same eigenvalues are quadratically small. Hence, if a diagonalization method is applied to such a matrix, then this property has an impact on the rate of convergence of the method and on the accuracy of the computed eigenvalues/eigenvectors (cf. [7, 14, 8]). These properties of an almost-diagonal matrix have several generalizations [6, 5, 7, 4, 9], especially to the pairs of almost-diagonal matrices. However, all these results make use of the absolute gaps in spectrum and are therefore less satisfactory in the case of close eigenvalues.

In this paper we derive appropriate estimates for a definite pair (H, M) of Hermitian matrices which are almost diagonal in the scaled sense, as defined in [1]. The obtained bounds depend on the relative gaps in spectrum, hence the new results are especially applicable when the eigenvalues of (H, M) cluster around the origin. The results are simplified provided H or M is diagonal. For example, if $M = I_n$ and $H = (h_{ij})$ is Hermitian with $h_{11} \geq \cdots \geq h_{nn}$, then for the nonincreasingly ordered eigenvalues of H we obtain

$$|h_{ii} - \lambda_i|/|\lambda_i| \leq (C/\gamma)\,\omega^2,\ 1 \leq i \leq n,\ \ C \ \text{ of order unity},$$

provided that $\omega < \gamma/4$. Here $\gamma = \min_{\lambda_i \neq \lambda_j} |\lambda_i - \lambda_j|/(|\lambda_i| + |\lambda_j|)$ and

$$\omega = \|D^{-1/2}\,(H - \mathrm{diag}\,(h_{11}, \ldots, h_{nn}))\,D^{-1/2}\|_2\,, \quad D = \mathrm{diag}(|h_{11}|, \ldots, |h_{nn}|).$$

Estimates of this type have already been derived for skew-Hermitian matrices [10] and for Hermitian positive definite matrix pairs [3]. The estimates presented in this

* Received by the editors December 12, 1994; accepted for publication (in revised form) by P. Van Dooren October 31, 1996.

http://www.siam.org/journals/simax/18-4/27847.html

† Department of Mathematics, University of Zagreb, Bijenička 30, 10000 Zagreb, Croatia (hari@math.hr). The research of this author was supported by Croatian Ministry of Science grant 1-01-252.

‡ Department of Computer Science, University of Colorado, Boulder, CO 80309-0430 (zlatko@cs.colorado.edu). The research of this author was supported by National Science Foundation grant ASC-9357812 and Department of Energy grant DE-FG03-94ER25215.

paper have nice applications in the quadratic convergence theory of scaled iterates by Hermitian Jacobi methods.

The paper is divided into three sections. In section 1 we introduce notation and present some known results on almost diagonal and on scaled diagonally dominant matrices. In section 2 we derive new estimates for the scaled diagonally dominant definite matrix pairs. Finally, in section 3 we apply the new results to a single Hermitian matrix and provide two examples.

1. A pair of almost-diagonal matrices. Consider the generalized eigenvalue problem

$$Hx = \lambda Mx, \quad x \neq \mathbf{0}, \tag{1.1}$$

with Hermitian H and Hermitian positive definite M, both of order n. Let the eigenvalues of the pair (H, M), that is, of the problem (1.1), be ordered nonincreasingly,

$$\lambda_1 = \cdots = \lambda_{s_1} > \lambda_{s_1+1} = \cdots = \lambda_{s_2} > \cdots > \lambda_{s_{p-1}+1} = \cdots = \lambda_{s_p}, \tag{1.2}$$

where $s_p = n$. Then for each $1 \leq i \leq p$, $n_i = s_i - s_{i-1}$, $(s_0 \stackrel{\text{def}}{=} 0)$ is the multiplicity of λ_{s_i}, respectively. For the pair (H, M) we assume

$$\frac{h_{11}}{m_{11}} \geq \frac{h_{22}}{m_{22}} \geq \cdots \geq \frac{h_{nn}}{m_{nn}}, \tag{1.3}$$

where $H = (h_{ij})$, $M = (m_{ij})$. To get such an arrangement one can apply to H and M the congruence transformation with a suitably chosen permutation matrix. According to the partition $n = n_1 + \cdots + n_p$ we define the block matrix partition

$$X = (x_{ij}) = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{p1} & \cdots & X_{pp} \end{bmatrix}, \qquad X_{ii} \in \mathbf{C}^{n_i \times n_i},\ 1 \leq i \leq p.$$

In order to move the block X_{ii} to the $(1,1)$ position we make use of the permutation matrices P_i, defined by

$$P_1 = I_n, \quad P_i = \left[e_{s_{i-1}+1}, \ldots, e_{s_i}, e_1, \ldots, e_{s_{i-1}}, e_{s_i+1}, \ldots\right], \quad 2 \leq i \leq p, \tag{1.4}$$

where $I_n = [e_1, \ldots, e_n]$ is the identity matrix. Then

$$P_i^* X P_i = \begin{bmatrix} \pi_i(X) & \tau_i(X) \\ \tau_i^c(X) & \pi_i^c(X) \end{bmatrix} \begin{matrix} \}n_i \\ \}n - n_i \end{matrix},$$

with

$$\pi_i(X) = X_{ii}, \qquad \tau_i(X) = \begin{bmatrix} X_{i1} & \ldots & X_{i,i-1} & X_{i,i+1} & \ldots \end{bmatrix},$$

$$\tau_i^c(X) = \begin{bmatrix} X_{1i} \\ \vdots \\ X_{i-1,i} \\ X_{i+1,i} \\ \vdots \end{bmatrix}, \quad \pi_i^c(X) = \begin{bmatrix} X_{11} & \cdots & X_{1,i-1} & X_{1,i+1} & \cdots \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ X_{i-1,1} & \cdots & X_{i-1,i-1} & X_{i-1,i+1} & \\ X_{i+1,1} & \cdots & X_{i+1,i-1} & X_{i+1,i+1} & \\ \vdots & \ldots & \ldots & \ldots & \ddots \end{bmatrix}.$$

Below, we use the following notation: $\boldsymbol{\Omega}(X) = X - \operatorname{diag}(X)$ is the off-diagonal part of X, $\pi(X) = \operatorname{diag}(X_{11}, \ldots, X_{pp})$ is the block-diagonal, and $\tau(X) = X - \pi(X)$ is the off-block-diagonal part of X. By $\|X\|_F$, $\|X\|_2$, and $\|X\|_\infty$ are denoted the Frobenius, the spectral (operator), and the infinity norm of X. As usual, $X^* = \overline{X}^T$.

1.1. Structure of an almost-diagonal pair. Let $\lambda(H,M)$ denote the spectrum of (H,M). Then for each $1 \le i \le p$

$$\delta_i = \min_{\substack{1\le j\le p\\ j\neq i}} \mid \lambda_{s_i} - \lambda_{s_j} \mid$$

is the (absolute) gap or separation of λ_{s_i} from $\lambda(H,M) \setminus \{\lambda_{s_i}\}$. The minimum separation in the spectrum $\lambda(H,M)$ is δ, where

$$\delta = \min_{1\le i\le p} \delta_i.$$

Let $M = \Delta_M M_S \Delta_M$ with $\Delta_M = [\mathrm{diag}\ (M)]^{1/2}$. The following result from [7, Theorem 3.1] reveals the special structure of an almost-diagonal pair (H,M):

(R1) If $\|\boldsymbol{\Omega}(\Delta_M^{-1} H \Delta_M^{-1} - \lambda_{s_i} M_S)\|_2 \le \frac{1}{3}\delta_i,\ 1 \le i \le p,$ then

$$\left\|\pi_i \left(\Delta_M^{-1} H \Delta_M^{-1} - \lambda_{s_i} M_S\right)\right\|_F \le \frac{3}{\delta_i} \left\|\tau_i \left(\Delta_M^{-1} H \Delta_M^{-1} - \lambda_{s_i} M_S\right)\right\|_F^2, \quad 1 \le i \le p.$$

In the case $M = I_n$ this result has the refinement [7], which provides information on the location of the eigenvalues of H:

(R2) If $\|\boldsymbol{\Omega}(H)\|_F \le \delta/3$, then

$$\|H_{ii} - \lambda_{s_i} I_{n_i}\|_F \le (1.32/\delta_i) \sum_{\substack{j=1\\ j\neq i}}^{p} \|H_{ij}\|_F^2, \ 1 \le i \le p.$$

Both results played essential roles in deriving sharp quadratic convergence estimates for appropriate Jacobi methods (see [5, 8, 11, 4]). Note that for tiny δ_i the appropriate bounds in (R1) and (R2) become large and therefore useless.

1.2. Pair of scaled diagonally dominant matrices. Here we recall the notion of *scaled diagonally dominant* matrices and matrix pairs from [1] and state some appropriate results. If $A = D + N$, where D is diagonal and N has zero diagonal, then $A = (a_{ij})$ is α-*diagonally dominant* with respect to a norm $\|\cdot\|$ if $\|N\| \le \alpha \min_{1\le i\le n} |a_{ii}|$, with $0 \le \alpha < 1$. Now, let $A = D + N$ with $|a_{ii}| = 1$, $1 \le i \le n$, and let Δ_1, Δ_2 be arbitrary nonsingular diagonal matrices. Then $B = \Delta_1 A \Delta_2$ is α-*scaled diagonally dominant* (α-s.d.d.) with respect to a given norm, if A is α-diagonally dominant with respect to that norm. If B is Hermitian, it is presumed that $\Delta_1 = \Delta_2$. Note that an α-s.d.d matrix has nonzero diagonal elements.

The pair (H,M) of Hermitian matrices is (α,β)-*scaled diagonally dominant definite*[1] ((α,β)-s.d.d.d.) with respect to a given norm if H is α-s.d.d., M is β-s.d.d., both with respect to that norm, and M is positive definite. If H is positive definite as well, (H,M) is (α,β)-s.d.d. positive definite ((α,β)-s.d.d.p.d.).

The spectral absolute value $[H^2]^{1/2}$ of a Hermitian matrix H, is denoted as in [15] by $\left|H\right|$. The standard absolute value of H is denoted by $|H|$, $|H| = (|h_{ij}|)$. The following result [15, Theorem 2.1] has nice applications for (η_H,η_M)-s.d.d.d. pairs.

(R3) Let (H,M) be a pair of Hermitian matrices such that M is positive definite. Let δH and δM be Hermitian matrices satisfying

$$\mid x^*\delta H x \mid \le \eta_H x^* \left|H\right|_M x,\ \eta_H < 1 \quad \text{and} \quad \mid x^*\delta M x \mid \le \eta_M x^* M x,\ \eta_M < 1$$

[1] This is a slight modification of the definition from [1].

for all $x \in \mathbf{C}^n$. Here $|H|_M = Z|Z^{-1}HZ^{-*}|Z^*$, where Z is any square matrix satisfying $M = ZZ^*$. If $\lambda_1 \geq \cdots \geq \lambda_n$ and $\lambda'_1 \geq \cdots \geq \lambda'_n$ are the eigenvalues of (H, M) and $(H + \delta H, M + \delta M)$, respectively, then for any $1 \leq i \leq n$, either $\lambda_i = \lambda'_i = 0$ or

$$\frac{1 - \eta_H}{1 + \eta_M} \leq \frac{\lambda'_i}{\lambda_i} \leq \frac{1 + \eta_H}{1 - \eta_M}.$$

We recall the result [2, Proposition 2.10], which refines [1, Proposition 2], and also (R3) for the pair $(\mathrm{diag}(H), I_n)$ with perturbation $(\Omega(H), 0)$, provided that H is positive definite.

(R4) If H is positive definite and $H_S = \Delta_H^{-1} H \Delta_H^{-1}$, $\Delta_H = [\mathrm{diag}(H)]^{1/2}$, then

$$1 - \|\boldsymbol{\Omega}(H_S)\|_2 \leq \lambda_{\min}(H_S) \leq \frac{\lambda_i}{h_{ii}} \leq \lambda_{\max}(H_S) \leq 1 + \|\boldsymbol{\Omega}(H_S)\|_2, \quad 1 \leq i \leq n,$$

where $\lambda_{\min}(H_S)$ and $\lambda_{\max}(H_S)$ are the smallest and the largest eigenvalues of H_S, respectively.

The results (R3) and (R4) will be used in sections 2 and 3, respectively.

2. Scaled almost-diagonal matrix pairs. Here we derive a result similar to (R1), but using scaled matrices and relative gaps in $\lambda(H, M)$. Relative gaps are defined in a number of ways (cf. [1, 15, 12, 10]). For technical reasons we use the following (relative) measure of eigenvalue separation.

DEFINITION 2.1. *Let λ_{s_i}, $1 \leq i \leq p$, be the eigenvalues of the pair (H, M), satisfying the condition* (1.2). *The* relative gap *of λ_{s_i} from $\lambda(H, M) \setminus \{\lambda_{s_i}\}$ is*

$$\gamma_i = \min_{\substack{1 \leq j \leq p \\ j \neq i}} \frac{|\, \lambda_{s_i} - \lambda_{s_j} \,|}{|\, \lambda_{s_i} \,| + |\, \lambda_{s_j} \,|}.$$

The minimum relative gap in $\lambda(H, M)$ is

$$\gamma = \min_{1 \leq i \leq p} \gamma_i.$$

Note that $\gamma_i \leq 1$, $1 \leq i \leq p$, and $\gamma_i = 1$ if either $\lambda_{s_i} = 0$ or λ_{s_i} is a single point of $\lambda(H, M)$ in $(-\infty, 0)$ or $(0, \infty)$.

The following theorem generalizes and improves [3, Theorem 2.24], which has been formulated for (α, α)-s.d.d.p.d. pairs.

THEOREM 2.2. *Let (H, M) be an (α, β)-s.d.d.d. pair satisfying the condition* (1.3). *Let $H = \Delta_H H_S \Delta_H$, $M = \Delta_M M_S \Delta_M$ with $\Delta_H = (|\mathrm{diag}\,(H)|)^{1/2}$, $\Delta_M = (\mathrm{diag}\,(M))^{1/2}$. If*

$$\frac{\alpha + \beta}{1 - \alpha} < \frac{1}{3}\,\gamma, \tag{2.1}$$

then for each $1 \leq i \leq p$ it holds that

(i) $$\|\pi_i \left(H_S - \lambda_{s_i} \Delta_H^{-1} M \Delta_H^{-1}\right) \|_F \leq \frac{4}{\gamma_i} \|\tau_i \left(H_S - \lambda_{s_i} \Delta_H^{-1} M \Delta_H^{-1}\right) \|_F^2,$$

(ii) $$\|\pi_i \left(H - \lambda_{s_i} M\right) \|_F \leq \frac{4}{\gamma_i} \|\tau_i \left(H \Delta_H^{-1} - \lambda_{s_i} M \Delta_H^{-1}\right) \|_F^2,$$

(iii) $$\|\pi_i \left(\lambda_{s_i}^{-1} \Delta_M^{-1} H \Delta_M^{-1} - M_S\right) \|_F \leq \frac{2}{\gamma_i} \|\tau_i \left(\lambda_{s_i}^{-1} \Delta_M^{-1} H \Delta_M^{-1} - M_S\right) \|_F^2,$$

(iv) $$\|\pi_i\left(\lambda_{s_i}^{-1}H - M\right)\|_F \leq \frac{2}{\gamma_i}\|\tau_i\left(\lambda_{s_i}^{-1}H\Delta_M^{-1} - M\Delta_M^{-1}\right)\|_F^2.$$

Proof. By the assumption we have

$$H_S = J_n + \boldsymbol{\Omega}(H_S), \quad \|\boldsymbol{\Omega}(H_S)\|_2 \leq \alpha < 1,$$
$$M_S = I_n + \boldsymbol{\Omega}(M_S), \quad \|\boldsymbol{\Omega}(M_S)\|_2 \leq \beta < 1,$$

where $J_n = I_k \bigoplus -I_{n-k}$ for some $0 \leq k \leq n$. Using substitutions $y = \Delta_H x$ and $z = \Delta_M x$ we obtain for arbitrary $x \in \mathbf{C}^n$

$$| x^*\boldsymbol{\Omega}(H)x | = | x^*\Delta_H\boldsymbol{\Omega}(H_S)\Delta_H x | = | y^*\boldsymbol{\Omega}(H_S)y | \leq \alpha y^*y = \alpha x^*\Delta_H^2 x, \tag{2.2}$$
$$| x^*\boldsymbol{\Omega}(M)x | = | x^*\Delta_M\boldsymbol{\Omega}(M_S)\Delta_M x | = | z^*\boldsymbol{\Omega}(M_S)z | \leq \beta z^*z = \beta x^*\Delta_M^2 x. \tag{2.3}$$

Consider (H, M) as the perturbed pair $\left(\Delta_H J_n \Delta_H, \Delta_M^2\right)$. Since

$$\left|\Delta_H J_n \Delta_H\right|_{\Delta_M^2} = \Delta_M\left|\Delta_M^{-1}\Delta_H J_n \Delta_H \Delta_M^{-1}\right|\Delta_M = \Delta_H^2,$$

the relations (2.2) and (2.3) imply that the result (R3) can be applied to the pair $\left(\Delta_H J_n \Delta_H, \Delta_M^2\right)$ with $(\boldsymbol{\Omega}(H), \boldsymbol{\Omega}(M))$ as perturbation. One obtains

$$\frac{1-\alpha}{1+\beta} \leq \frac{\lambda_{s_i}}{h_{jj}/m_{jj}} \leq \frac{1+\alpha}{1-\beta}, \quad j \in \mathcal{S}_i,\ 1 \leq i \leq p,$$

where $\mathcal{S}_i = \{\, j\,;\, s_{i-1}+1 \leq j \leq s_i\},\ 1 \leq i \leq p$. Hence

$$\frac{|\, h_{jj}/m_{jj} - \lambda_{s_i}\,|}{|\,\lambda_{s_i}\,|} \leq \frac{\alpha+\beta}{1-\alpha}, \quad j \in \mathcal{S}_i,\ 1 \leq i \leq p. \tag{2.4}$$

Now, consider the intervals

$$\mathcal{D}_i = \left\{\xi\ :\ \frac{|\,\xi - \lambda_{s_i}\,|}{|\,\lambda_{s_i}\,|} < \frac{\gamma_i}{3}\right\}, \quad 1 \leq i \leq p.$$

Since for $\xi \in \mathcal{D}_k$, $k \neq i$,

$$\begin{aligned}\frac{|\xi - \lambda_{s_i}|}{|\lambda_{s_i}|} &\geq \frac{|\lambda_{s_i} - \lambda_{s_k}|}{|\lambda_{s_i}| + |\lambda_{s_k}|}\left(1 + \frac{|\lambda_{s_k}|}{|\lambda_{s_i}|}\right) - \frac{|\xi - \lambda_{s_k}|}{|\lambda_{s_k}|}\frac{|\lambda_{s_k}|}{|\lambda_{s_i}|} \\ &\geq \max\{\gamma_i, \gamma_k\} + \frac{2}{3}\frac{|\lambda_{s_k}|}{|\lambda_{s_i}|}\max\{\gamma_i, \gamma_k\} \geq \max\{\gamma_i, \gamma_k\},\end{aligned} \tag{2.5}$$

we see that $\xi \notin \mathcal{D}_i$. Thus, $\mathcal{D}_i$, $1 \leq i \leq p$, are mutually disjoint. If (2.1) holds, then the relation (2.4) implies $h_{jj}/m_{jj} \in \mathcal{D}_i$, $j \in \mathcal{S}_i$, $1 \leq i \leq p$.

Let $i \in \{1, \ldots, p\}$ be fixed. Consider $C_i = H - \lambda_{s_i} M$. If P_i is as in (1.4), then

$$P_i^T C_i P_i = \begin{bmatrix} \pi_i(C_i) & \tau_i(C_i) \\ \tau_i^c(C_i) & \pi_i^c(C_i) \end{bmatrix}$$

has rank $n - n_i$ and the rank argument (cf. [7, 17]) implies

$$\pi_i(C_i) = \tau_i(C_i)\left[\pi_i^c(C_i)\right]^{-1}\tau_i^c(C_i), \tag{2.6}$$

provided that $\pi_i^c(C_i)$ is nonsingular.

Now we prove the assertions (i)–(iv) of the theorem.

(i) We use $\pi_i(\Delta_H^{-1})$ and $\pi_i^c(\Delta_H^{-1})$ as scaling matrices for modifying (2.6). We obtain

$$\pi_i(\tilde{C}_i) = \tau_i(\tilde{C}_i)\left[\pi_i^c(\tilde{C}_i)\right]^{-1}\tau_i^c(\tilde{C}_i) \tag{2.7}$$

with $\tilde{C}_i = \Delta_H^{-1}C_i\Delta_H^{-1}$. Note that $\pi_i^c(\tilde{C}_i) = \pi_i^c(\Delta_H^{-1})\pi_i^c(C_i)\pi_i^c(\Delta_H^{-1})$ is nonsingular if and only if $\pi_i^c(C_i)$ is nonsingular; hence (2.6) holds if and only if (2.7) holds. Let μ be the smallest by modulus eigenvalue of $\pi_i^c(\tilde{C}_i)$. If we show

$$|\,\mu\,| > \frac{\gamma_i}{4}, \tag{2.8}$$

then (2.7) holds with $\|\left[\pi_i^c(\tilde{C}_i)\right]^{-1}\|_2 < 4/\gamma_i$. Then, applying the Frobenius matrix norm to (2.7) and using $\tau_i^c(\tilde{C}_i) = \tau_i(\tilde{C}_i)^*$, we obtain the assertion (i) of the theorem. Let $\pi_i^c(\tilde{C}_i)w = \mu w$, $w \neq 0$. This can be written as $\pi_i^c(C_i)w' = \mu\pi_i^c(\Delta_H^2)w'$, $w' = \pi_i^c(\Delta_H^{-1})w \neq 0$. Hence

$$\left[\pi_i^c\left(\Delta_H(J_n - \mu I_n)\Delta_H\right) + \boldsymbol{\Omega}\left(\pi_i^c(H)\right)\right]w' = \lambda_{s_i}\left[\pi_i^c(\Delta_M^2) + \boldsymbol{\Omega}\left(\pi_i^c(M)\right)\right]w',$$

implying that λ_{s_i} is an eigenvalue of the pair $\left(\pi_i^c(H - \mu\Delta_H^2)\,,\,\pi_i^c(M)\right)$, which depends on μ. If we let $y = \pi_i^c(\Delta_H)x$ and $z = \pi_i^c(\Delta_M)x$, we have for arbitrary $x \in \mathbf{C}^{n-n_i}$

$$\begin{aligned}
|\,x^*\boldsymbol{\Omega}\left(\pi_i^c(H)\right)x\,| &= |\,y^*\boldsymbol{\Omega}\left(\pi_i^c(H_S)\right)y\,| \leq \|\boldsymbol{\Omega}\left(\pi_i^c(H_S)\right)\|_2 y^*y \leq \|\boldsymbol{\Omega}(H_S)\|_2 y^*y \\
&\leq \alpha y^*y = \alpha x^*\pi_i^c(\Delta_H^2)x = \frac{\alpha}{1-|\mu|}x^*\left(1-|\mu|\right)\pi_i^c(\Delta_H^2)x \\
&\leq \frac{\alpha}{1-|\mu|}x^*\pi_i^c\left(\Delta_H|J_n - \mu I_n|\Delta_H\right)x \\
&= \frac{\alpha}{1-|\mu|}x^*\left|\pi_i^c\left(\Delta_H(J_n - \mu I_n)\Delta_H\right)\right|_{\pi_i^c(\Delta_M^2)}x, \\
|\,x^*\boldsymbol{\Omega}\left(\pi_i^c(M)\right)x\,| &= |\,z^*\boldsymbol{\Omega}\left(\pi_i^c(M_S)\right)z\,| \leq \|\boldsymbol{\Omega}\left(\pi_i^c(M_S)\right)\|_2 z^*z \leq \|\boldsymbol{\Omega}(M_S)\|_2 z^*z \\
&\leq \beta z^*z = \beta x^*\pi_i^c(\Delta_M^2)x.
\end{aligned}$$

Here we have used the inequality $(1-|\mu|)I_n \leq |J_n - \mu I_n|$ and the fact that the spectral norm of a submatrix is not larger than the norm of the whole matrix. The latest two relations imply that the result (R3) can be applied to $\left(\pi_i^c(\Delta_H(J_n - \mu I_n)\Delta_H), \pi_i^c(\Delta_M^2)\right)$ with $(\boldsymbol{\Omega}(\pi_i^c(H)), \boldsymbol{\Omega}(\pi_i^c(M)))$ as perturbation, provided that $|\mu| < 1 - \alpha$. If this is the case we obtain (since λ_{s_i} is an eigenvalue of $\left(\pi_i^c(H - \mu\Delta_H^2)\,,\,\pi_i^c(M)\right)$)

$$\frac{1 - \dfrac{\alpha}{1-|\mu|}}{1+\beta} \leq \frac{\lambda_{s_i}}{\dfrac{h_{jj}}{m_{jj}}(1-\tilde{\mu})} \leq \frac{1 + \dfrac{\alpha}{1-|\mu|}}{1-\beta} \quad \text{for some } j \in \bigcup_{k\neq i}\mathcal{S}_k,$$

with $\tilde{\mu} = \mu\,\mathrm{sign}(h_{jj})$. Using elementary calculus one obtains

$$\frac{|h_{jj}/m_{jj} - \lambda_{s_i}|}{|\lambda_{s_i}|} \leq \frac{\alpha + \beta + |\mu|}{1 - \alpha - |\mu|} \quad \text{for some } j \in \bigcup_{k\neq i}\mathcal{S}_k. \tag{2.9}$$

By (2.5) we have $|\,h_{jj}/m_{jj} - \lambda_{s_i}\,|/|\lambda_{s_i}| > \gamma_i$. Hence (2.9) and (2.1) imply

$$\gamma_i < \frac{\alpha + \beta + |\mu|}{1 - \alpha - |\mu|} < \frac{(1-\alpha)\frac{1}{3}\gamma + |\mu|}{1 - \alpha - |\mu|},$$

and we obtain $|\mu| > (1-\alpha)\frac{3\gamma_i-\gamma}{3(\gamma_i+1)}$. Since $\beta \geq 0$ the assumption (2.1) implies $1-\alpha > 3/(3+\gamma)$. We also have $3\gamma_i - \gamma \geq 2\gamma_i$. Hence $|\mu| > \gamma_i/4$.

If $|\mu| \geq 1-\alpha$ one obtains straightforwardly $|\mu| > 3/(3+\gamma) \geq 3/4 > \gamma_i/4$. Since i is arbitrary the proof of the assertion (i) is completed.

(ii) Here we scale only the right-hand side of (2.6) with $\pi_i^c(\Delta_H^{-1})$ and obtain

$$\pi_i(C_i) = \tau_i(C_i\Delta_H^{-1}) \left[\pi_i^c(\tilde{C}_i)\right]^{-1} \tau_i^c(\Delta_H^{-1}C_i). \tag{2.10}$$

By the relation (2.8) we know that $\|\left[\pi_i^c(\tilde{C}_i)\right]^{-1}\|_2 \leq 4/\gamma_i$. We also have $[\tau_i(C_i\Delta_H^{-1})]^* = \tau_i^c(\Delta_H^{-1}C_i)$. Hence, (ii) follows by applying the Frobenius norm to (2.10).

(iii) We scale (2.6) with $\pi_i(\Delta_M^{-1})$ and $\pi_i^c(\Delta_M^{-1})$ and multiply both sides of the obtained equality by $\lambda_{s_i}^{-1}$. We obtain

$$\pi_i(C_i') = \tau_i(C_i')\left[\pi_i^c(C_i')\right]^{-1}\tau_i^c(C_i'),$$

where $C_i' = \lambda_{s_i}^{-1}\Delta_M^{-1}H\Delta_M^{-1} - M_S$. Let ν be an eigenvalue of $\pi_i^c(C_i')$ such that $1/\mid\nu\mid = \|[\pi_i^c(C_i')]^{-1}\|_2$. Then $\pi_i^c(C_i')v = \nu v$ for some $v \neq 0$. With $v' = \pi_i^c(\Delta_M^{-1})v$ we have

$$\pi_i^c(H)v' = \lambda_{s_i}\left[(\nu+1)\pi_i^c\left(\Delta_M^2\right) + \boldsymbol{\Omega}\left(\pi_i^c(M)\right)\right]v', \quad v' \neq 0.$$

Thus λ_{s_i} is an eigenvalue of the pair $\left(\pi_i^c(H)\,,\,\pi_i^c(M+\nu\Delta_M^2)\right)$. As above, one can show that for an arbitrary $x \in \mathbf{C}^{n-n_i}$

$$\begin{aligned}
&\mid x^*\boldsymbol{\Omega}\left(\pi_i^c(H)\right)x \mid \leq \alpha x^*\pi_i^c(\Delta_H^2)x = \alpha x^*\left|\pi_i^c\left(\Delta_H J_n \Delta_H\right)\right|_{\pi_i^c((1+\nu)\Delta_M^2)}x,\\
&\mid x^*\boldsymbol{\Omega}\left(\pi_i^c(M)\right)x \mid \leq \frac{\beta}{1+\nu}x^*\pi_i^c(\,(1+\nu)\Delta_M^2)x.
\end{aligned}$$

If $\nu > \beta - 1$ one can apply (R3) to the pair $\left(\pi_i^c(\Delta_H J_n \Delta_H), (1+\nu)\pi_i^c(\Delta_M^2)\right)$ with $(\boldsymbol{\Omega}(\pi_i^c(H))\ ,\ \boldsymbol{\Omega}(\pi_i^c(M)))$ as perturbation, to obtain

$$\frac{1-\alpha}{1+\dfrac{\beta}{\nu+1}} < \frac{\lambda_{s_i}}{\dfrac{h_{jj}}{(\nu+1)m_{jj}}} < \frac{1+\alpha}{1-\dfrac{\beta}{\nu+1}} \quad \text{for some} \quad j \in \bigcup_{k\neq i}\mathcal{S}_k.$$

The latest relation together with the condition (2.1) implies

$$\frac{|h_{jj}/m_{jj} - \lambda_{s_i}|}{|\lambda_{s_i}|} < \frac{\alpha+\beta+\nu}{1-\alpha} < \frac{1}{3}\gamma + \frac{\nu}{1-\alpha}.$$

By (2.5), we obtain

$$\gamma_i - \frac{1}{3}\gamma < \frac{\nu}{1-\alpha}, \tag{2.11}$$

implying $\nu > 0$. Since $1/(1-\alpha) < 1+\gamma/3$, we have $\nu/(1-\alpha) < (1+\gamma/3)\nu$. This inequality together with (2.11) implies $|\nu| > \gamma_i/2$.

If $\nu < \beta - 1$, we have $\mid\nu\mid > 1-\beta > 1-\gamma/3 > 2/3 > \gamma_i/2$. Hence, in any case, $\mid\nu\mid > \gamma_i/2$. This proves (iii), since i is arbitrary.

(iv) The proof follows the lines of the proof of (ii), except that Δ_M and the assertion (iii) are used instead of Δ_H and the assertion (i), respectively. ∎

Note that any unitarily invariant matrix norm can be applied to both sides of (2.7). In the case of the spectral norm one obtains (cf. Example 3.3 and Corollary 3.2) the claim on the first page of the paper.

If both matrices H and M are positive definite, we can further improve the latest result.

COROLLARY 2.3. *Suppose the pair (H, M) is (α, β)-s.d.d.p.d. and let H_S, Δ_H, M_S, and Δ_M be as in Theorem* 2.2.

(a) *If in* (2.1) *the denumerator $1-\alpha$ is replaced by $1-\beta$, then the constants* 4 *and* 2 *in the assertions* (i), (ii) *and* (iii), (iv), *respectively, interchange their places.*

(b) *If in* (2.1) *the denumerator $1-\alpha$ is replaced by $1-\max\{\alpha,\beta\}$, then all the assertions of Theorem* 2.2 *hold with the same constant* 2 *on the right-hand sides.*

Proof. (a) The pair (M, H) is (β, α)-s.d.d.p.d. with eigenvalues $\lambda_{s_p}^{-1} = \cdots = \lambda_{s_{p-1}+1}^{-1} > \cdots > \lambda_{s_1}^{-1} = \cdots = \lambda_1^{-1}$ and with the appropriate gaps $\gamma_j(M,H) = \gamma_{p+1-j}(H,M)$, $1 \le j \le p$. Hence the assertion (iii) of Theorem 2.2 for the pair (M, H) takes the form

$$\|\pi_j\Big((\lambda_{s_j}^{-1})^{-1}\Delta_H^{-1}M\Delta_H^{-1} - H_S\Big)\| \le \frac{2}{\gamma_j}\|\tau_j\Big((\lambda_{s_j}^{-1})^{-1}\Delta_H^{-1}M\Delta_H^{-1} - H_S\Big)\|_F^2, 1 \le j \le p$$

and this is (i) with 4 replaced by 2. In the same way, starting with the assertion (i) of Theorem 2.2, one can prove that (iii) holds with the constant 4 instead of 2. The same argument applies to the assertions (ii) and (iv) as well.

(b) The proof follows directly from the assertions (iii) and (iv) of Theorem 2.2 and the assertion (a) of this corollary. ∎

3. Some applications. When M or H is diagonal we can use the congruence transformation with Δ_M or Δ_H, to reduce it to I_n or J_n. This leads us to the pairs $(\Delta_M^{-1}H\Delta_M^{-1}, I_n)$ and $(\Delta_H^{-1}M\Delta_H^{-1}, J)$, that is, to the estimates for Hermitian and J-Hermitian matrices. We shall consider in detail only the case of Hermitian matrices. The estimates for J-Hermitian matrices can be obtained in a similar way.

Scaled almost-diagonal Hermitian matrix. Let λ_j, $1 \le j \le n$, be the eigenvalues of the Hermitian matrix H ordered nonincreasingly, as indicated by the relation (1.2). We further assume the nonincreasing ordering of the diagonal elements of H. This assumption corresponds to the relation (1.3). The sets $\mathcal{S}_i$ are defined as above. Note also the definition of π and τ just before section 1.1.

COROLLARY 3.1. *Let H be an α-s.d.d. Hermitian matrix and let $H = \Delta_H H_S \Delta_H$, $\Delta_H = [|\mathrm{diag}(H)|]^{1/2}$. If $\alpha < \gamma/(\gamma+3)$, then*

(i) $$\sum_{j\in\mathcal{S}_i}\left|1-\frac{\lambda_{s_i}}{h_{jj}}\right|^2 + \|\boldsymbol{\Omega}(\pi_i(H_S))\|_F^2 \le \frac{16}{\gamma_i^2}\|\tau_i(H_S)\|_F^4, \quad 1 \le i \le p,$$

(ii) $$\sum_{j=1}^{n}\left|1-\frac{\lambda_j}{h_{jj}}\right|^2 + \|\boldsymbol{\Omega}(\pi(H_S))\|_F^2 \le \frac{8}{\gamma^2}\|\tau(H_S)\|_F^4,$$

(iii) $$\sum_{j\in\mathcal{S}_i} |\, h_{jj} - \lambda_{s_i}\,|^2 + \|\boldsymbol{\Omega}(\pi_i(H))\|_F^2 \le \frac{4}{\gamma_i^2}\min\left\{4\|\tau_i(H\Delta_H^{-1})\|_F^4, \frac{\|\tau_i(H)\|_F^4}{\lambda_{s_i}^2}\right\}, \quad 1 \le i \le p.$$

(iv) $$\sum_{j=1}^{n} |\, h_{jj} - \lambda_j\,|^2 + \|\boldsymbol{\Omega}(\pi(H))\|_F^2 \le \frac{2}{\gamma^2}\min\left\{8\|\tau(H\Delta_H^{-1})\|_F^4, \frac{\|\tau(H)\|_F^4}{\min_{1\le i\le p}\lambda_{s_i}^2}\right\}.$$

Proof. (i) The proof follows directly from Theorem 2.2 by inserting $M = M_S = \Delta_M = I_n$. (ii) If we sum the left- and right-hand sides of the inequality in (i) and use $\gamma_i \geq \gamma$, $\|\tau_i(H_S)\|^2 \leq \|\tau(H_S)\|^2/2$ for $1 \leq i \leq p$, we obtain (ii). (iii) The proof follows by combining the assertions (ii) and (iii) (or (ii) and (iv)) of Theorem 2.2. (iv) The proof is similar to the proof of (ii). We use (iii) and the fact that $\|\tau_i(H\Delta_H^{-1})\| \leq \|\tau(H\Delta_H^{-1})\|$, $1 \leq i \leq p$. ∎

If λ_{s_i} is simple, then the terms $\|\boldsymbol{\Omega}(\pi_i(H_S))\|_F^2$ and $\|\boldsymbol{\Omega}(\pi_i(H))\|_F^2$ in (i) and (iii) are omitted. Similarly, if all the eigenvalues of H are simple, then $\|\boldsymbol{\Omega}(\pi(H_S))\|_F^2$ and $\|\boldsymbol{\Omega}(\pi(H))\|_F^2$ in (ii) and (iv) are omitted. We have included in (iii) and (iv) the term containing the eigenvalue(s) λ_{s_i} because knowledge of rough lower bound(s) for λ_{s_i} may considerably improve the bound (see Example 3.3 below). If H is, in addition, positive definite, we can further sharpen the latest result.

COROLLARY 3.2. *Let H be an α-s.d.d. Hermitian positive definite matrix and let $H = \Delta_H H_S \Delta_H$, $\Delta_H = [\mathrm{diag}(H)]^{1/2}$. If $\alpha < \gamma/3$, then*

$$\text{(i)}\quad \sum_{j\in\mathcal{S}_i}\left|1-\frac{\lambda_{s_i}}{h_{jj}}\right|^2 + \|\boldsymbol{\Omega}(\pi_i(H_S))\|_F^2 \leq \frac{4}{\gamma_i^2}\|\tau_i(H_S)\|_F^4, \quad 1 \leq i \leq p,$$

$$\text{(ii)}\quad \sum_{j=1}^{n}\left|1-\frac{\lambda_j}{h_{jj}}\right|^2 + \|\boldsymbol{\Omega}(\pi(H_S))\|_F^2 \leq \frac{2}{\gamma^2}\|\tau(H_S)\|_F^4.$$

If $\alpha < \gamma/(\gamma+3)$, then

$$\text{(iii)}\quad \sum_{j\in\mathcal{S}_i} |\, h_{jj}-\lambda_{s_i}\,|^2 + \|\boldsymbol{\Omega}(\pi_i(H))\|_F^2 \leq \frac{4}{\gamma_i^2}\min\left\{\|\tau_i(H\Delta_H^{-1})\|_F^4, \frac{\|\tau_i(H)\|_F^4}{\lambda_{s_i}^2}\right\}, \quad 1 \leq i \leq p,$$

$$\text{(iv)}\quad \sum_{j=1}^{n} |\, h_{jj}-\lambda_j\,|^2 + \|\boldsymbol{\Omega}(\pi(H))\|_F^2 \leq \frac{2}{\gamma^2}\min\left\{2\|\tau(H\Delta_H^{-1})\|_F^4, \frac{\|\tau(H)\|_F^4}{\min_{1\leq i\leq p}\lambda_{s_i}^2}\right\}.$$

Proof. Consider the pairs (H, I_n) and (I_n, H) and apply already derived estimates (or equivalently, use Corollaries 2.3 and 3.1). ∎

In the following example we shall comparatively investigate to what relative accuracy the diagonals of H reveal the eigenvalues of H if the results (R2), (R4), and Corollary 3.2 are applied.

Example 3.3. Let (cf. [3, Example 2.8])

$$H = \begin{bmatrix} 1 & 10^{-8} & 10^{-8} \\ 10^{-8} & 10^{-6} & 10^{-11} \\ 10^{-8} & 10^{-11} & 10^{-8} \end{bmatrix}, \quad M = I_3.$$

Since $\|\boldsymbol{\Omega}(H)\|_2 \leq \|\boldsymbol{\Omega}(H)\|_\infty = 2\cdot 10^{-8}$ the perturbation theorem for the eigenvalues of symmetric matrices (or simply the Gerschgorin theorem) implies $|\,\lambda_i - h_{ii}\,| \leq 2\cdot 10^{-8}$. Hence the absolute gaps can be bounded from below using the formula $\delta_i \geq \min_{j,\,j\neq i}|h_{ii}-h_{jj}| - 2\cdot(2\cdot 10^{-8})$. One obtains $\delta_1 > 9.999\cdot 10^{-1}$, $\delta_2 = \delta_3 = \delta > 9.500\cdot 10^{-7}$. Therefore, the result (R2) implies

$$|\,\lambda_1 - h_{11}\,| < 1.321\cdot 10^{-16}, \quad \max\{|\,\lambda_2 - h_{22}\,|, |\,\lambda_3 - h_{33}\,|\} < 1.390\cdot 10^{-10}. \tag{3.1}$$

Let $r_i = |\,\lambda_i - h_{ii}\,|/\lambda_i$, $1 \leq i \leq 3$. Since $\lambda_1 > h_{11} - 1.321\cdot 10^{-16}$ and $\lambda_j > h_{jj} - 1.390\cdot 10^{-10}$, $2 \leq j \leq 3$, the relation (3.1) implies

$$r_1 < 1.322\cdot 10^{-16}, \quad r_2 < 1.31\cdot 10^{-4}, \quad r_3 < 1.410\cdot 10^{-2}.$$

Thus, only two (four) significant digits of λ_3 (λ_2) can be revealed.

Let us now apply the result (R4) to H. Since

$$H_S = \Delta_H^{-1} H \Delta_H^{-1} = \begin{bmatrix} 1 & 10^{-5} & 10^{-4} \\ 10^{-5} & 1 & 10^{-4} \\ 10^{-4} & 10^{-4} & 1 \end{bmatrix},$$

we have $\|\boldsymbol{\Omega}(H_S)\|_2 \le \|\boldsymbol{\Omega}(H_S)\|_\infty = 2 \cdot 10^{-4} \equiv \alpha$. Hence $1 - \alpha < \lambda_i / h_{ii} < 1 + \alpha$, $1 \le i \le 3$, implying

$$\frac{|\, \lambda_i - h_{ii} \,|}{\lambda_i} < \frac{2 \cdot 10^{-4}}{1 - 2 \cdot 10^{-4}} < 2.001 \cdot 10^{-4}, \quad 1 \le i \le 3.$$

The result (R4) enables us to reveal two more digits in λ_3 than (R2). However, for the largest eigenvalue λ_1 the result (R2) yields better information than (R4).

Consider now the estimates based on Gerschgorin disks[2] for DHD^{-1}, where D is a suitably chosen diagonal matrix (see [16, sections 2.14, 2.15] or [13, section IV 2.2]). In particular, if $D = \mathrm{diag}(d_1, 1, 1)$, then the smallest value of d_1 for which the disk around 1 remains isolated from other disks is approximately $1.00000100001 \cdot 10^{-8}$. The Gerschgorin theorem applied to DHD^{-1} yields $r_1 < 2.0000021 \cdot 10^{-16}$. If $D = \mathrm{diag}(1, d_2, 1)$, then the optimal value for d_2 is approximately $1.02040826962 \cdot 10^{-5}$, and it yields $r_2 < 1.02143 \cdot 10^{-7}$. Finally, if $D = \mathrm{diag}(1, 1, d_3)$, then the optimal value for d_3 is again $1.02040826962 \cdot 10^{-5}$, and it yields $r_3 < 1.02144 \cdot 10^{-5}$. We see that this technique yields better bounds than the result (R4) and, except for r_1, much better bounds than the result (R2).

Let us apply the estimates of Corollary 3.2 to H. The basic estimates $|\, \lambda_i - h_{ii} \,| \le 2 \cdot 10^{-8}$, $1 \le i \le 3$, yield the following lower bounds for the relative gaps in the spectrum of H:

$$\gamma_1 = \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2} > \frac{h_{11} - h_{22} - |\, \lambda_1 - h_{11} \,| - |\, \lambda_2 - h_{22} \,|}{h_{11} + h_{22} + |\, \lambda_1 - h_{11} \,| + |\, \lambda_2 - h_{22} \,|} > 0.9999979,$$

$$\gamma = \gamma_2 = \gamma_3 = \frac{\lambda_2 - \lambda_3}{\lambda_2 + \lambda_3} > \frac{h_{22} - h_{33} - |\, \lambda_2 - h_{22} \,| - |\, \lambda_3 - h_{33} \,|}{h_{22} + h_{33} + |\, \lambda_2 - h_{22} \,| + |\, \lambda_3 - h_{33} \,|} > 0.9047619.$$

Consider first the estimates based on Corollary 3.2 (iii). Using the above bounds for γ_i we obtain

$$|\, h_{11} - \lambda_1 \,| \le \frac{2}{\gamma_1} \min \left\{ (10^{-10} + 10^{-8}), \frac{10^{-16} + 10^{-16}}{1 - 2 \cdot 10^{-8}} \right\} < 4.00001 \cdot 10^{-16},$$

$$|\, h_{22} - \lambda_2 \,| \le \frac{2}{\gamma_2} \min \left\{ 10^{-16} + 10^{-14}, \frac{10^{-16} + 10^{-22}}{10^{-6} - 2 \cdot 10^{-8}} \right\} < 2.23264 \cdot 10^{-14},$$

$$|\, h_{33} - \lambda_3 \,| \le \frac{2}{\gamma_3} \left(10^{-16} + 10^{-16} \right) < 4.42106 \cdot 10^{-16}.$$

All these bounds are pretty sharp and, except for $|\, h_{11} - \lambda_1 \,|$, much better than those based on the result (R2). Let us now apply Corollary 3.2 (i) to H. Since $(2/\gamma_1)(1 + 10^{-2}) \cdot 10^{-8} < 2.021 \cdot 10^{-8}$, we have

$$\left| \frac{\lambda_1}{h_{11}} - 1 \right| < 2.021 \cdot 10^{-8} \ \text{ implying } \ r_1 < \frac{2.021 \cdot 10^{-8}}{1 - 2.021 \cdot 10^{-8}} < 2.022 \cdot 10^{-8}.$$

[2] The authors would like to thank one of the referees for reminding them of this technique.

In a similar way we obtain

$$r_2 < 2.23263 \cdot 10^{-8}, \quad r_3 < 4.422 \cdot 10^{-8}.$$

The latest two bounds might have been improved slightly had $\gamma_2 = \gamma_3$ been computed via the estimates (3.1). In that case we would have obtained $r_2 < 2.063 \cdot 10^{-8}$ and $r_3 < 4.084 \cdot 10^{-8}$. Since[3] $(\lambda_3 - h_{33})/\lambda_3 \approx -2.01008085848 \cdot 10^{-8}$ the bound for the smallest eigenvalue is realistic. In general, the first (third) assertion of Corollary 3.2 will yield better bounds for the smallest (largest) eigenvalues of a positive definite matrix.

Although in Example 3.3 the technique with Gerschgorin circles gave better bounds than the result (R4), the latter is much more convenient on some occasions (e.g., verifying convergence of the Jacobi eigenreduction method).

The next example shows how good the new estimates are when H is two by two.

Example 3.4. Let

$$H = \begin{bmatrix} a & b \\ \bar{b} & c \end{bmatrix}, \quad a \geq c,$$

be a two by two Hermitian matrix. Since

$$\begin{aligned} \lambda_1 &= a + \mid b \mid \tan \varphi, \\ \lambda_2 &= c - \mid b \mid \tan \varphi, \end{aligned} \qquad \tan 2\varphi = \frac{2 \mid b \mid}{a - c}, \quad \mid \varphi \mid \leq \frac{\pi}{4},$$

we can calculate exactly

$$1 - \frac{\lambda_1}{a} = -\frac{\mid b \mid}{a} \tan \varphi = -\frac{2 \mid b \mid^2}{a\left[a - c + \sqrt{(a-c)^2 + 4 \mid b \mid^2}\right]} \quad \text{if} \quad a \neq 0,$$

$$1 - \frac{\lambda_2}{c} = \frac{\mid b \mid}{c} \tan \varphi = \frac{2 \mid b \mid^2}{c\left[a - c + \sqrt{(a-c)^2 + 4 \mid b \mid^2}\right]} \quad \text{if} \quad c \neq 0.$$

Suppose first that H is definite; that is, a and c are of the same sign and $ac > \mid b \mid^2$. The latest relations show that

$$\left| 1 - \frac{\lambda_1}{a} \right| \leq \frac{1}{\mid a \mid} \frac{\mid b \mid^2}{a - c}, \qquad \left| 1 - \frac{\lambda_2}{c} \right| \leq \frac{1}{\mid c \mid} \frac{\mid b \mid^2}{a - c}.$$

On the other hand, since

$$\gamma_1 = \gamma_2 = \gamma = \frac{\mid \lambda_1 - \lambda_2 \mid}{\mid \lambda_1 \mid + \mid \lambda_2 \mid} = \frac{a - c + 2 \mid b \mid \tan \varphi}{\mid a + c \mid},$$

the requirement $\|\boldsymbol{\Omega}(H_S)\|_2 \equiv \alpha < \gamma/3$ of Corollary 3.2 reads

$$3\frac{\mid b \mid}{\sqrt{ac}} < \frac{a - c}{\mid a + c \mid} + 2\frac{\mid b \mid \tan \varphi}{\mid a + c \mid}.$$

If the latest inequality holds, Corollary 3.2(i) implies

$$\max\left\{ \left| 1 - \frac{\lambda_1}{a} \right|, \left| 1 - \frac{\lambda_2}{c} \right| \right\} \leq 2\frac{\mid a + c \mid}{a - c} \frac{\mid b \mid^2}{ac} = 2\frac{\mid a + c \mid}{ac} \frac{\mid b \mid^2}{a - c}.$$

[3] We have computed the eigenvalues of H with accuracy 10^{-80} by the interactive system MATHEMATICA.

Since

$$\max\left\{\frac{1}{|\,a\,|}\,,\ \frac{1}{|\,c\,|}\right\} \le \frac{1}{2}\left\{2\frac{|\,a+c\,|}{a-c}\right\},$$

the bound of Corollary 3.2(i) is at least twice as large as the actual bound. Note also that the condition $\alpha < \gamma/3$ is superfluous.

Yet, consider the case $a > 0 > c$. Now H is indefinite, and therefore $\gamma_1 = \gamma_2 = \gamma = 1$. Hence the requirement $\alpha < \gamma/(\gamma+3)$ of Corollary 3.1 reads $4 \mid b \mid < \sqrt{a \mid c \mid}$. So, under that condition, Corollary 3.1(i) implies

$$\max\left\{\left|1-\frac{\lambda_1}{a}\right|\,,\ \left|1-\frac{\lambda_2}{c}\right|\right\} \le \frac{4}{1}\frac{|\,b\,|^2}{a\mid c\mid}.$$

Since

$$\max\left\{\frac{|\,b\,|^2}{a^2+a\mid c\mid}\,,\ \frac{|\,b\,|^2}{c^2+a\mid c\mid}\right\} \le \frac{|\,b\,|^2}{a\mid c\mid},$$

the bound in Corollary 3.1(i) can be sharp. However, the requirement $\alpha < \gamma/(\gamma+3)$ appears to be superfluous.

Remark 3.5. Our estimates can easily be applied to a class of non-Hermitian matrices which are Hermitian up to diagonal scaling. More precisely, we refer to $H \in \mathbf{C}^{n\times n}$ as *hidden Hermitian* if $H = D_1 A D_2$ for some nonsingular real diagonal matrices D_1, D_2 and Hermitian A. Since the eigenvalue problem $Hx = \lambda x$ is equivalent to the generalized eigenvalue problem for the Hermitian pair $(A, D_1^{-1}D_2^{-1})$, our results can be applied provided that either A or D_1D_2 is positive definite.

Appendix. Here we prove relation (2.9). Since $1-\tilde{\mu} > 0$, the latest relation is equivalent to

$$-l \equiv -\frac{-\tilde{\mu}+\frac{1-\tilde{\mu}}{1-|\mu|}\alpha+\beta}{1-\tilde{\mu}+\frac{1-\tilde{\mu}}{1-|\mu|}\alpha} \le \frac{\frac{h_{jj}}{m_{jj}}-\lambda_{s_i}}{\lambda_{s_i}} \le \frac{\tilde{\mu}+\frac{1-\tilde{\mu}}{1-|\mu|}\alpha+\beta}{1-\tilde{\mu}-\frac{1-\tilde{\mu}}{1-|\mu|}\alpha} \equiv u.$$

If $\tilde{\mu} = |\mu|$, then

$$l = \frac{-|\mu|+\alpha+\beta}{1-|\mu|+\alpha} \le \frac{\alpha+\beta+|\mu|}{1-\alpha-|\mu|} = u.$$

If $\tilde{\mu} = -|\mu|$, then we have

$$l = \frac{\alpha+\frac{1-|\mu|}{1+|\mu|}(\beta+|\mu|)}{1-|\mu|+\alpha} \le \frac{\alpha+\beta+|\mu|}{1-\alpha-|\mu|}$$

and

$$u = \frac{t+\beta}{1-t} \quad \text{with} \quad t = \frac{1+|\mu|}{1-|\mu|}\alpha - |\mu|.$$

Since the condition $\alpha/(1-|\mu|) < 1$ is equivalent to $t < \alpha+|\mu|$, we obtain

$$u \le \frac{\alpha+\beta+|\mu|}{1-\alpha-|\mu|}.$$

Hence, in both cases, we have proven

$$\frac{|h_{jj}/m_{jj}-\lambda_{s_i}|}{|\lambda_{s_i}|} \le \frac{\alpha+\beta+|\mu|}{1-\alpha-|\mu|} \quad \text{for some } j \in \bigcup_{k\neq i}\mathcal{S}_k.$$

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] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.

[3] Z. Drmač, *Computing the Singular and the Generalized Singular Values*, Ph.D. thesis, Fernuniversität Hagen, Germany, 1994.

[4] Z. Drmač and V. Hari, *On the quadratic convergence of the J-symmetric Jacobi method*, Numer. Math., 64 (1993), pp. 147–180.

[5] V. Hari, *On Cyclic Jacobi Methods for the Positive Definite Generalized Eigenvalue Problem*, Ph.D. thesis, Fernuniversität Hagen, Germany, 1984.

[6] V. Hari, *On almost diagonal square matrices with multiple singular values*, Rad. Mat., 4 (1988), pp. 209–225.

[7] V. Hari, *On pairs of almost diagonal matrices*, Linear Algebra Appl., 148 (1991), pp. 193–223.

[8] V. Hari, *On sharp quadratic convergence bounds for the serial Jacobi methods*, Numer. Math., 60 (1991), pp. 375–406.

[9] V. Hari and N. H. Rhee, *A matrix pair of an almost diagonal skew-symmetric matrix and a symmetric positive definite matrix*, Linear Algebra Appl., 148 (1993), pp. 82–117.

[10] E. Pietzsch, *Genaue Eigenwertberechnung nichtsingulärer schiefsymmetrischer Matrizen mit einem Jacobi–änlichen Verfahren*, Ph.D. thesis, Fernuniversität Hagen, Germany, 1993.

[11] I. Slapničar and V. Hari, *On the quadratic convergence of the Falk–Langemeyer method*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 84–114.

[12] I. Slapničar and K. Veselić, *Perturbations of the eigenprojections of a factorised Hermitian matrix*, Linear Algebra Appl., 218 (1995), pp. 273–280.

[13] G. W. Stewart and J. Guang Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[14] H. P. M. van Kempen, *On quadratic convergence of the special cyclic Jacobi method*, Numer. Math., 9 (1966), pp. 19–22.

[15] K. Veselić and I. Slapničar, *Floating-point perturbations of Hermitian matrices*, Linear Algebra Appl., 195 (1993), pp. 81–116.

[16] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Springer-Verlag, Berlin, Heidelberg, New York, 1965.

[17] J. H. Wilkinson, *Almost diagonal matrices with multiple or close eigenvalues*, Linear Algebra Appl., 1 (1968), pp. 1–12.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 4, pp. 1013–1034, October 1997

014

ON COMPUTING AN EIGENVECTOR OF A TRIDIAGONAL MATRIX. PART I: BASIC RESULTS*

K. V. FERNANDO†

Abstract. We consider the solution of the homogeneous equation $(J - \lambda I)x = 0$, where J is a tridiagonal matrix, λ is a known eigenvalue, and x is the unknown eigenvector corresponding to λ. Since the system is underdetermined, x could be obtained by setting $x_k = 1$ and solving for the rest of the elements of x. This method is not entirely new, and it can be traced back to the times of Cauchy [*Oeuvres Complétes* (II^e Série), Vol. 9, Gauthier–Villars, Paris, 1841]. In 1958, Wilkinson demonstrated that, in finite-precision arithmetic, the computed x is highly sensitive to the choice of k; the traditional practice of setting $k = 1$ or $k = n$ can lead to disastrous results. We develop algorithms to find optimal k which require an LDU and a UDL factorization (where L is lower bidiagonal, D is diagonal, and U is upper bidiagonal) of $J - \lambda I$ and are based on the theory developed by Fernando [*On a Classical Method for Computing Eigenvectors*, Numerical Algorithms Group Ltd, Oxford, 1995] for general matrices. We have also discovered new formulae (valid also for more general Hessenberg matrices) for the determinant of $J - \tau I$, which give better numerical results when the shifted matrix is nearly singular. These formulae could be used to compute eigenvalues (or to improve the accuracy of known estimates) based on standard zero finders such as Newton and Laguerre methods. The accuracy of the computed eigenvalues is crucial for obtaining small residuals for the computed eigenvectors. The algorithms for solving eigenproblems are embarrassingly parallel and hence suitable for modern architectures.

Key words. eigenvalues, eigenvectors, perturbation analysis, tridiagonal matrices, deflation, inverse iteration

AMS subject classifications. 15A18, 15A23, 15A24

PII. S0895479895294484

1. Introduction and summary.

It appears that there are not many algorithms to compute eigenvectors of a matrix once the eigenvalues are determined. In fact, inverse iteration seems to be the only mainstream algorithm in the repertoire of readily available software. Although inverse iteration [13] is a very powerful tool, there are many well-known shortcomings. Some of them include the following:

1. the use of random vectors which defies deterministic analysis;
2. the need for reorthogonalization of computed vectors using Gram–Schmidt or otherwise which can be expensive, especially on parallel platforms;
3. the inability to obtain orthogonal vectors when the eigenvalues are clustered;
4. the excessive reliance on heuristics.

The method we are proposing is not entirely new; the basic idea has been in existence for nearly two centuries since the times of Cauchy [4] and is based on the definition of an eigenvector. However, in the late 1950s this method went into disrepute because Wilkinson [17] uncovered a fundamental obstacle to this classical approach; more details are given later in this section. The essentials of the method are based on the solution of the set of homogeneous equations $(F - \lambda I)\boldsymbol{x} = 0$, where F is a general square matrix and $\boldsymbol{x}$ is the unknown eigenvector corresponding to the known eigenvalue λ. The same approach has been used for the generalized eigenvalue problem

* Received by the editors November 8, 1995; accepted for publication (in revised form) by A. Greenbaum October 31, 1996. A preliminary version of this paper appeared in *Proc. of the Third International Congress on Industrial and Applied Mathematics*, Zeitschrift für Angewandte Mathematik und Mechanik, Hamburg, 1995.

http://www.siam.org/journals/simax/18-4/29448.html

† Numerical Algorithms Group Ltd, Wilkinson House, Jordan Hill, Oxford OX2 8DR, UK (vince@nag.co.uk).

$(F - \lambda G)\boldsymbol{x} = 0$. Since these are underdetermined systems of equations, at least one equation in the system is redundant. If the kth equation is redundant, then one may assume that x_k, the kth element of $\boldsymbol{x}$, is unity and solve the rest of the equations. It has been the normal practice to assume that the superfluous equation is the nth or the first. The Holzer method in vibration analysis [12], which has been in existence since the turn of this century, also follows this tradition.

Is it possible to drop any one of the n equations and obtain a good approximation to an eigenvector? This question has been asked by many in the past; see Fernando [7], which also gives a detailed historical account. Note that if the matrix F is dense, then the nth equation does not have a particular significance since any equation can be given the ordinal count n after a trivial permutation of the equations. However, for tridiagonal and other structured matrices, reordering the equations may destroy the structure. Thus, dropping the first or the nth equation has become entrenched for tridiagonal matrices and pencils.

In a pioneering paper [17], Wilkinson showed that not all equations of a homogeneous system are equal; some have a higher degree of redundancy than others. The removal of an equation which is less redundant than others can lead to disastrous results. This is not a problem which afflicts only large matrices; it can happen even if $n = 2$. We recall the example of Wilkinson (see section 51 of [19]):

$$F = \begin{pmatrix} 0.713263 & 0.000984 \\ 0.000984 & 0.121665 \end{pmatrix}, \quad \lambda_1 = 0.71326463, \quad \lambda_2 = 0.12166336.$$

Wilkinson took the approximation $\lambda \approx \tau = 0.713265$ which has an approximate error of 0.36×10^{-6}. By omitting the second equation of the nearly homogeneous system $(F - \tau I)\boldsymbol{z} \approx 0$, he obtained the computed eigenvector

$$\boldsymbol{z}^t = (1, 0.002033)\,. \tag{1}$$

When he dropped the first equation, the eigenvector was

$$\boldsymbol{z}^t = (1, 0.00166329)\,. \tag{2}$$

The eigenvector given by (1) is accurate only to 3 decimal places while the eigenvector (2) is accurate to 8 decimal places.

In [7], Fernando[1] studied in depth the theoretical issues concerning the choice of k for general dense matrices, and the results of Wilkinson were reconfirmed and extended. The basic idea is to compute the diagonal entries of the matrix M, which is obtained by elementwise reciprocation of the inverse of the matrix $(F - \lambda I)^t$. Thus, for the Wilkinson example

$$M = \begin{pmatrix} -0.00000036 & -0.00021844 \\ -0.00021844 & -9.3047491 \end{pmatrix}.$$

The diagonal element of M with the smallest magnitude points to the equation which should be dropped. Thus, $k = 1$ is the optimal choice.

One of the main objectives of this paper is to develop algorithms to compute the diagonal elements of M for tridiagonal matrices. This is based on an LDU and a UDL factorization of the shifted tridiagonal matrix $J - \tau I$ which give all possible

[1] The technical report [7] is not a prerequisite to follow this paper; however, for a deeper understanding, it should be consulted.

burn at both ends (BABE) factorizations. This report was motivated by an article by Henrici [11] who seems to be the first person to use BABE factorizations. We study the accuracy of BABE factorizations of tridiagonal matrices elsewhere [6].

We also encounter a variable defined by Babuška [2], which is present in his error analysis of tridiagonal equation solvers. In our notation, this variable is μ_k, the kth diagonal element of the matrix M.

Suppose that the LDU factorization of $J-\tau I$ exists. It can be shown that the last pivot d_n of the LDU factorization of $J-\tau I$ is zero if the shift τ is an exact eigenvalue of J. However, in finite precision arithmetic d_n can be huge even if the shift is very accurate. We explain this phenomenon of the nonvanishing d_n. The algorithm for finding the optimal kth equation which should be dropped from a nearly homogeneous system was discovered while this problem was studied.

One of the highlights of this paper is the formulae we have discovered for the determinant of a shifted tridiagonal matrix in terms of the leading/trailing principal minors and the $\mu_k(\tau)$ (the kth diagonal element of M). In fact, our results are valid for a wider class of problems defined by Hessenberg matrices; thus we formulate our problem in terms of the Hessenberg matrix H. The formulae are

$$\det(H-\tau I) = \mu_k(\tau)\,\det(H-\tau I)_{1:k-1}\,\det(H-\tau I)_{k+1:n},\quad 1\le k\le n.$$

The significance of this result is due to the fact that if the shifted matrix is singular then it is often possible to obtain, in floating-point arithmetic, a zero (or a tiny) determinant provided we choose k such that $|\mu_k(\tau)|$ is minimal. In exact arithmetic, if the matrix $H-\tau I$ is singular then all $\mu_k(\tau)$ values are zero; but in floating-point arithmetic, the convergence of $\mu_k(\tau)$ to zero can be uneven for different values of k as indicated by the 2×2 example. Since many eigenvalue solvers are based on the premise of a vanishing determinant when the shift is an exact eigenvalue, these formulae provide a means to achieve that objective in floating-point arithmetic. We have already applied Newton and Laguerre zero-finding techniques to the determinants given by the new formulae; we hope to discuss these experiments elsewhere.

In floating-point arithmetic, the determinant of a matrix is not a good indicator of the singularity of that matrix (see section 2.7.3 of Golub and Van Loan [10]). However, by choosing k such that $|\mu_k|$ is minimal it is possible to estimate nearly singular determinants more accurately.

The kth diagonal entry μ_k of M can be interpreted as the perturbation required to make the matrix $J-\tau I$ singular when τ is not an exact eigenvalue of J; that is,

$$\det(J-\tau I-\mu_k \boldsymbol{e}_k\boldsymbol{e}_k^*)=0,$$

where $\boldsymbol{e}_k$ is the unit vector with unity at the kth entry and zeros elsewhere. However, if the tridiagonal matrix has zero diagonals, which is the case if the matrix is related to a bidiagonal singular value decomposition (SVD) problem, perturbation of the diagonal elements is contraindicated. We remove this difficulty by perturbing a pair of off-diagonal elements instead of a diagonal element.

This paper is organized as follows. In section 2, the notation and the preliminaries are established. The basic theory concerning near homogeneous systems of equations is developed in section 3. The formulae for computing the diagonal values of M which indicate the levels of redundancy are derived in section 4. The algorithms for computing eigenvectors, once the optimal k is known, are covered in section 5. In section 6, diagonal perturbations are avoided by transferring the disturbance to a pair of off-diagonal elements. The quality of the computed eigenvectors is assessed

in section 7. Finally, in section 8, the mystery of the nonvanishing pivot d_n with an ideal shift is investigated and solved.

The algorithms for computing eigenvectors are embarrassingly parallel, and they are suitable for modern architectures. Elsewhere we treat the problems associated with eigenvectors corresponding to clustered eigenvalues.

2. Notation and preliminaries. Scalars are denoted by lowercase Greek and Roman characters. Eigenvalues are shown as λ_i's and singular values as σ_i's.

Vectors are denoted by bold Roman characters such as $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$. The unit vector $\boldsymbol{e}_k$ has a one in the kth element and zeros elsewhere. The kth element of $\boldsymbol{x}$ is x_k, the transpose of $\boldsymbol{x}$ is $\boldsymbol{x}^t$, and the complex conjugate transpose of $\boldsymbol{x}$ is $\boldsymbol{x}^*$.

Matrices are shown as uppercase Roman characters. The submatrix of F containing the rows from i to j and columns from k to l is indicated by $F_{i:j,k:l}$, which is consistent with Fortran and MATLAB notations; the contracted form $F_{i:j}$ is used for $F_{i:j,i:j}$. We use the shorthand notation $F_{[k,l]}$ for the matrix obtained by removing kth row and the lth column of the matrix F. Similarly, $\boldsymbol{y}_{[k]}$ is the vector we get if the kth element is omitted.

We use J to denote an $n \times n$ (possibly complex) tridiagonal matrix. The nonzero elements of the ith row of J are $(c_{i-1},\ a_i,\ b_i)$ with a_i centered on the diagonal (except that c_0 does not exist in the first row and b_n does not exist in the nth row). We assume that the tridiagonal matrix is unreduced; that is, none of the off-diagonal elements b_i and c_i are zero.

It is presumed that the reader knows the LDU factorization of a tridiagonal matrix where L is lower bidiagonal, U is upper bidiagonal, and D is diagonal. The diagonal values of L and U are all unity.

We consider the LDU factorization of the unreduced tridiagonal matrix $J - \tau I$ where the scalar τ is a shift. Here we use the notation $L(\tau)$, $D(\tau)$, and $U(\tau)$ for the factors, to emphasize their dependence on τ. The following recursion for computing the pivots $d_i(\tau)$ (which are the diagonal values of $D(\tau)$) is well known:

$$d_i(\tau) = a_i - \tau - b_{i-1}c_{i-1}/d_{i-1}(\tau), \quad i = 2, \ldots, n \ \text{ with } \ d_1(\tau) = a_1 - \tau. \tag{3}$$

The $(i+1, i)$ element of $L(\tau)$ is $c_i/d_i(\tau)$, and the $(i, i+1)$ element of $U(\tau)$ is $b_i/d_i(\tau)$. We often suppress the argument τ from $d_i(\tau)$ and similar variables.

If J is a Hermitian matrix (or a matrix which can be made Hermitian via a diagonal similarity transformation—that is, $b_i c_i > 0$ for $i = 1, \ldots, n-1$), then the number of positive (negative) d_i gives the number of eigenvalues of J which are greater (less) than τ. Thus the recursion (3) can be found in the inner loop of most bisection algorithms for finding eigenvalues of symmetric tridiagonal matrices. See Golub and Van Loan [10, section 8.4.1] and Kahan [14]. Let the number of negative d_i (the inertia count) be $\nu(\tau)$.

If the UDL factorization of $J - \tau I$ is defined as $\tilde{U}(\tau)\Delta(\tau)\tilde{L}(\tau)$, then the pivots $\delta_i(\tau)$ (the diagonal elements of $\Delta(\tau)$) are given by

$$\delta_i(\tau) = a_i - \tau - b_i c_i/\delta_{i+1}(\tau), \quad i = n-1, \ldots, 1 \ \text{ with } \ \delta_n(\tau) = a_n - \tau. \tag{4}$$

The $(i+1, i)$ element of $\tilde{L}(\tau)$ is $c_i/\delta_{i+1}(\tau)$, and the $(i, i+1)$ element of $\tilde{U}(\tau)$ is $b_i/\delta_{i+1}(\tau)$.

We use the following simple but useful result extensively, without explicitly invoking the lemma.

LEMMA 2.1. *If L is unit lower triangular and $L\boldsymbol{x} = \boldsymbol{e}_n$, then $\boldsymbol{x} = \boldsymbol{e}_n$. Similarly, if U is unit upper triangular and $U\boldsymbol{x} = \boldsymbol{e}_1$, then $\boldsymbol{x} = \boldsymbol{e}_1$.*

Most of our developments are based on *LDU* and *UDL* factorization of submatrices $J - \tau I$. However, it is possible to use orthogonal factorizations, instead of Gaussian factorizations, without any major problems. Since the cost of orthogonal factorizations is high, many would prefer to use Gaussian factorizations with nearest neighbor pivoting to minimize forward errors. However, since QR/QL factorizations are the basis of the QR/QL algorithms for computation of eigenvalues and eigenvectors, the formulae for computing μ_k are paramount in the understanding of the QR/QL algorithms. This topic is studied elsewhere.

We denote the diagonal elements of the matrix R of the QR factorization of $(J - \tau I)_{1:k}$ by $\{r_1, r_2, \ldots, r_{k-1}, \hat{r}_k\}$. Note that the corresponding values for $(J - \tau I)_{1:k+1}$ are $\{r_1, \ldots, r_k, \hat{r}_{k+1}\}$, where $r_k = \sqrt{\hat{r}_k^2 + c_k^2}$ and $\hat{r}_1 = a_1 - \tau$.

Similarly, let $QL = (J - \tau I)_{k:n}$ be the QL factorization of $(J - \tau I)_{k:n}$, where the diagonal elements of L are designated as $\{\hat{l}_k, l_{k+1}, \ldots, l_n\}$ with $l_k = \sqrt{\hat{l}_k^2 + b_{k-1}^2}$ and $\hat{l}_n = a_n - \tau$.

The following lemma which gives the ratios of principal minors is easy to verify.

LEMMA 2.2.

$$\frac{\det(J - \tau I)_{1:k-1}}{\det(J - \tau I)_{1:k}} = \frac{\hat{r}_{k-1}}{r_{k-1}\hat{r}_k} = \frac{\cos\alpha_{k-1}}{\hat{r}_k} = \frac{1}{d_k},$$

$$\frac{\det(J - \tau I)_{k+1:n}}{\det(J - \tau I)_{k:n}} = \frac{\hat{l}_{k+1}}{l_{k+1}\hat{l}_k} = \frac{\cos\beta_{k+1}}{\hat{l}_k} = \frac{1}{\delta_k},$$

$$\begin{pmatrix} \cos\alpha_k & e^{\imath\theta}\sin\alpha_k \\ -e^{-\imath\theta}\sin\alpha_k & \cos\alpha_k \end{pmatrix} \begin{pmatrix} \hat{r}_k \\ c_k \end{pmatrix} = \begin{pmatrix} r_k \\ 0 \end{pmatrix},$$

$$\begin{pmatrix} \cos\beta_k & -e^{-\imath\gamma}\sin\beta_k \\ e^{\imath\gamma}\sin\beta_k & \cos\beta_k \end{pmatrix} \begin{pmatrix} b_{k-1} \\ \hat{l}_k \end{pmatrix} = \begin{pmatrix} 0 \\ l_k \end{pmatrix},$$

where $\imath^2 = -1$.

If nearest neighbor pivoting of rows is incorporated, then the following formulae may be used. We use the prefix P in *LDU* and *UDL* to indicate row pivoting. See section 47, Chapter 5 of Wilkinson [19] for further details. The proof of the lemma is left as an exercise for the reader.

LEMMA 2.3. *Let* $\hat{d}_{k-1}$ *be the last pivot in the PLDU factorization of* $(J - \tau I)_{1:k-1}$ *with nearest neighbor pivoting and, similarly,* $\hat{\delta}_{k+1}$ *be the last pivot of the PUDL factorization* $(J - \tau I)_{k+1:n}$. *If no pivoting takes place at the kth step in the computation of* $\hat{d}_k$ *then*

$$\frac{\det(J - \tau I)_{1:k-1}}{\det(J - \tau I)_{1:k}} = \frac{1}{\hat{d}_k}.$$

Similarly, if pivoting is not present in the evaluation of $\hat{\delta}_k$ *at that step,*

$$\frac{\det(J - \tau I)_{k+1:n}}{\det(J - \tau I)_{k:n}} = \frac{1}{\hat{\delta}_k}.$$

However, if pivoting is present at these steps, then

$$\frac{\det(J-\tau I)_{1:k-1}}{\det(J-\tau I)_{1:k}} = -\frac{\hat{d}_{k-1}}{c_{k-1}\hat{d}_k},$$

$$\frac{\det(J-\tau I)_{k+1:n}}{\det(J-\tau I)_{k:n}} = -\frac{\hat{\delta}_{k+1}}{b_k\hat{\delta}_k}.$$

3. Solution of a nearly homogeneous system. Suppose that an eigenvalue, λ, of the (dense) matrix F is known exactly; the eigenvector $\boldsymbol{x}$ corresponding to λ is given by the homogeneous system of equations and is given by

$$(F-\tau I)\boldsymbol{x} = 0 \text{ with } \tau = \lambda. \tag{5}$$

This system can be solved by assuming an arbitrary nonzero value for x_k for a particular k, $k = 1, \ldots, n$. This is a reasonable assumption since not all elements of $\boldsymbol{x}$ can be zero. Thus, the rest of the solution is given by

$$\begin{pmatrix} F_{1:k-1}-\tau I & F_{1:k-1,k+1:n} \\ F_{k+1:n,1:k-1} & F_{k+1:n}-\tau I \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_{1:k-1} \\ \boldsymbol{x}_{k+1:n} \end{pmatrix} = -x_k \begin{pmatrix} F_{1:k-1,k} \\ F_{k+1:n,k} \end{pmatrix}. \tag{6}$$

The kth equation of (5) is of the form

$$(F-\tau I)_{k,1:n}\boldsymbol{x} = 0. \tag{7}$$

However, in practice, eigenvalues are not known exactly and hence the shift τ is not an exact eigenvalue of F. Thus, equation (7) will not be satisfied exactly even if the solution of (6) is known exactly. That is, instead of (7), we get

$$(F-\tau I)_{k,1:n}\boldsymbol{x}(\tau) = \mu_k(\tau), \tag{8}$$

where $\mu_k(\tau)$ is the residual. Thus, instead of (5), we are effectively solving the system of equations

$$(F-\tau I)\boldsymbol{y}(k,\tau) = \mu_k(\tau)\boldsymbol{e}_k,$$

where we have changed the notation from $\boldsymbol{x}$ to $\boldsymbol{y}(k,\tau)$ to emphasize that we are now solving for an approximate eigenvector $\boldsymbol{y}(k,\tau)$ corresponding to the approximate eigenvalue τ. From now on we often suppress the arguments k and τ of $\boldsymbol{y}(k,\tau)$ and similar vectors to avoid cluttering.

We now specialize our results for the tridiagonal matrix J for which a fundamental simplification is possible. The submatrices $J_{1:k-1,k+1:n}$ and $J_{k+1:n,1:k-1}$ are null matrices and thus (6) gives two decoupled systems of equations

$$(J-\tau I)_{1:k-1}\boldsymbol{y}_{1:k-1} = -c_{k-1}y_k\boldsymbol{e}_{k-1} \tag{9}$$

and

$$(J-\tau I)_{k+1:n}\boldsymbol{y}_{k+1:n} = -b_k y_k \boldsymbol{e}_1. \tag{10}$$

The residual equation (8) for the tridiagonal J is

$$c_{k-1}y_{k-1} + (a_k-\tau)y_k + b_k y_{k+1} = \mu_k(\tau). \tag{11}$$

LEMMA 3.1. *If $(J-\tau I)_{1:k-1}$ and $(J-\tau I)_{k+1:n}$ are nonsingular, then the equation*

$$(J-\tau I)\boldsymbol{y} = \mu_k(\tau)\boldsymbol{e}_k$$

has a nonzero solution $\boldsymbol{y}$ if and only if y_k is not zero.

Proof. From (9), $\boldsymbol{y}_{1:k-1}$ is zero if and only if y_k is zero since $(J-\tau I)_{1:k-1}$ is nonsingular. Similarly from (10), $\boldsymbol{y}_{k+1,n}$ is zero if and only if y_k is zero. Thus $\boldsymbol{y}$ is zero if and only if y_k is zero. ∎

The previous result indicates that y_k can be taken as a nonzero value for the solution of $\boldsymbol{y}$ whenever the two leading/trailing principal submatrices $(J-\tau I)_{1:k-1}$ and $(J-\tau I)_{k+1:n}$ are nonsingular. We now show that eigenvectors can have zero elements, but there are certain restrictions.

LEMMA 3.2. *Let $\boldsymbol{x}$ be an eigenvector of the unreduced tridiagonal matrix J corresponding to the eigenvalue λ. Then no two contiguous elements of (x_{i-1}, x_i, x_{i+1}) can be zero, where $1 \le i \le n$, with the extended end conditions $x_0 = x_{n+1} = 0$ (which we assume for notational convenience). Furthermore, x_{i-1} and x_{i+1} cannot be both zero if $a_i - \lambda \neq 0$.*

Proof. The proof is by contradiction. The ith equation of $(J-\lambda I)\boldsymbol{x} = 0$ is

$$c_{i-1}x_{i-1} + (a_i - \lambda)x_i + b_i x_{i+1} = 0.$$

If x_{i-1} and x_i are zero, then x_{i+1} is zero since b_i is nonzero. Thus, $x_j = 0$ for $j < i-1$ and $j > i+1$. This violates the property that any eigenvector has at least one nonzero element. A similar contradiction manifests itself if x_i and x_{i+1} are zero.

To prove the second part, assume that x_{i-1} and x_{i+1} are zero. Then $x_i = 0$ since $a_i - \lambda \neq 0$; in that case all the elements of $\boldsymbol{x}$ become zero. ∎

Remark. The first part of the lemma shows that $x_1 \neq 0$ and $x_n \neq 0$. According to the second part, $x_2 \neq 0$ if $a_1 \neq \lambda$ and $x_{n-1} \neq 0$ if $a_n \neq \lambda$.

THEOREM 3.3. *Let $(J-\tau I)_{1:k-1}$ and $(J-\tau I)_{k+1,n}$ be nonsingular and*

$$(J-\tau I)\boldsymbol{z}(k,\tau) = \mu_k(\tau)\boldsymbol{e}_k \quad \text{with} \quad z_k = 1. \tag{12}$$

Then the perturbed matrix $J - \mu_k(\tau)\boldsymbol{e}_k\boldsymbol{e}_k^* - \tau I$ is singular. Furthermore, τ is an exact eigenvalue of the perturbed matrix $J - \mu_k(\tau)\boldsymbol{e}_k\boldsymbol{e}_k^*$ and $\boldsymbol{z}$ is the corresponding eigenvector.

Proof. Since $z_k = 1$, (12) can be written in the form

$$(J-\tau I)\boldsymbol{z} = \mu_k(\tau)\boldsymbol{e}_k\boldsymbol{e}_k^*\boldsymbol{z}$$

and thus

$$[J - \mu_k(\tau)\boldsymbol{e}_k\boldsymbol{e}_k^* - \tau I]\boldsymbol{z} = 0.$$

Since $\boldsymbol{z}$ is nonzero (from Lemma 3.1), $J - \mu_k(\tau)\boldsymbol{e}_k\boldsymbol{e}_k^* - \tau I$ is singular. ∎

The preceding theorem suggests that τ should be chosen such that $\mu_k(\tau)$ is as tiny as possible if accurate eigenvectors are required. That is,

$$\boldsymbol{z}(k,\tau) \to \boldsymbol{x}, \quad \tau \to \lambda \quad \text{as} \quad \mu_k(\tau) \to 0$$

for any k provided $\boldsymbol{x}$ is unique (up to a scalar multiple).

The next three results are stated without a proof.

COROLLARY 3.4. *If $\mu_k(\tau)$ is zero for a fixed k, then $\mu_i(\tau)$ is zero for all i.*

COROLLARY 3.5. *If τ is an exact eigenvalue of the matrix J, then $\mu_i(\tau)$ is zero for all i, $i = 1, \ldots, n$.*

COROLLARY 3.6. *If J is real, then*

$$(J - \bar{\tau})\bar{\boldsymbol{z}}(k, \bar{\tau}) = \bar{\mu}_k(\bar{\tau})\boldsymbol{e}_k \quad \textit{with} \quad \bar{z}_k = 1.$$

Remark. If an eigenvalue has a complex conjugate partner, the μ_k have to be computed only once for both eigenvalues.

The following formula may be used for computing the residual $\mu_k(\tau)$ for a particular k. The essential idea, which is also valid for dense matrices, can be traced to Sherman and Morrison [15]. See Fernando [7].

THEOREM 3.7. *Let $(J - \tau I)_{1:k-1}$ and $(J - \tau I)_{k+1,n}$ be nonsingular and*

$$(J - \tau I)\boldsymbol{z}(k, \tau) = \mu_k(\tau)\boldsymbol{e}_k \ \textit{ with } \ z_k = 1.$$

Then $\mu_k(\tau) = 1/\boldsymbol{e}_k^*(J - \tau I)^{-1}\boldsymbol{e}_k$.

Proof. $\boldsymbol{z}$ is given by $\boldsymbol{z} = \mu_k(\tau)(J - \tau I)^{-1}\boldsymbol{e}_k$. Multiplying both sides of the above equation by $\boldsymbol{e}_k^*$ we arrive at the stated result. ∎

Remark. Thus, $\mu_k(\tau)$ is the reciprocal of the kth diagonal entry of the inverse of $J - \tau I$.

The following algorithm may be used for computing the approximate eigenvector $\boldsymbol{y}$ and the residual $\mu_k(\tau)$ for a particular k.

ALGORITHM 1 (the basic method).

1. *Choose an index k*
2. *Set τ to an estimate of an eigenvalue of J*
3. *Set $y_k = 1$*
4. *If $k \neq 1$, solve the equations $(J - \tau I)_{1:k-1}\boldsymbol{y}_{1:k-1} = -c_{k-1}\boldsymbol{e}_{k-1}$ for $\boldsymbol{y}_{1:k-1}$*
5. *If $k \neq n$, solve the equations $(J - \tau I)_{k+1:n}\boldsymbol{y}_{k+1:n} = -b_k\boldsymbol{e}_1$ for $\boldsymbol{y}_{k+1:n}$*
6. *Compute $\mu_k = c_{k-1}y_{k-1} + (a_k - \tau)y_k + b_k y_{k+1}$.*

The easiest way to solve the system of equations in step 4 of Algorithm 1 is to use the *LDU* factorization without pivoting. However, nearest neighbor pivoting can be used to avoid forward errors, in which case the U matrix becomes triangular with three upper diagonals. The QR factorization is a more expensive alternative which also creates a matrix R with three upper diagonals.

Similarly, for the solution of the set of equations in step 5, the *UDL* factorization is the obvious choice. If nearest neighbor pivoting is used, then L is tridiagonal and triangular. The most suitable orthogonal factorization for the solution is the QL.

So far we have not answered an important issue: how to find the index k such that the residual $|\mu_k(\tau)|$ is the smallest without computing the inverse of $(J - \tau I)^{-1}$ or equation (11) for every k.

4. Formulae for the residual $\boldsymbol{\mu_k(\tau)}$. It is possible to derive many interesting formulae to compute the residual $\mu_k(\tau)$.

THEOREM 4.1. *Let $(J - \tau I)_{1:k-1}$ and $(J - \tau I)_{k+1,n}$ be nonsingular and*

$$(J - \tau I)\boldsymbol{z}(k, \tau) = \mu_k(\tau)\boldsymbol{e}_k \ \textit{ with } \ z_k = 1. \tag{13}$$

Then $\mu_k(\tau)$ is given by

$$\mu_k(\tau) = (a_k - \tau) - b_{k-1}c_{k-1}\frac{\det(J - \tau I)_{1:k-2}}{\det(J - \tau I)_{1:k-1}} - b_k c_k \frac{\det(J - \tau I)_{k+2:n}}{\det(J - \tau I)_{k+1:n}}.$$

Proof. The first step is to expand the determinant of $J - \tau I - \mu_k(\tau)\boldsymbol{e}_k\boldsymbol{e}_k^*$ along the kth column. Note that

$$\det\{J - \tau I - \mu_k(\tau)\boldsymbol{e}_k\boldsymbol{e}_k^*\}_{[k-1,k]} = c_{k-1}\det(J - \tau I)_{1:k-2}\det(J - \tau I)_{k+1:n},$$

$$\det\{J - \tau I - \mu_k(\tau)\boldsymbol{e}_k\boldsymbol{e}_k^*\}_{[k,k]} = \det(J - \tau I)_{1:k-1}\det(J - \tau I)_{k+1:n},$$

$$\det\{J - \tau I - \mu_k(\tau)\boldsymbol{e}_k\boldsymbol{e}_k^*\}_{[k+1,k]} = b_k\det(J - \tau I)_{1:k-1}\det(J - \tau I)_{k+2:n}.$$

Then,

$$\begin{aligned}\det\{J - \tau I - \mu_k(\tau)\boldsymbol{e}_k\boldsymbol{e}_k^*\} = \{a_k - \tau - \mu_k(\tau)\}&\det(J - \tau I)_{1:k-1}\det(J - \tau I)_{k+1:n}\\ -b_{k-1}c_{k-1}&\det(J - \tau I)_{1:k-2}\det(J - \tau I)_{k+1:n}\\ -b_kc_k&\det(J - \tau I)_{1:k-1}\det(J - \tau I)_{k+2:n}.\end{aligned}$$

However, the determinant of $J-\tau I-\mu_k(\tau)\boldsymbol{e}_k\boldsymbol{e}_k^*$ is zero (see Theorem 3.3) which leads to the stated result. ∎

It is easy to establish three more formulae for $\mu_k(\tau)$.

COROLLARY 4.2.

$$\mu_k(\tau) = \frac{\det(J - \tau I)_{1:k}}{\det(J - \tau I)_{1:k-1}} - b_kc_k\frac{\det(J - \tau I)_{k+2:n}}{\det(J - \tau I)_{k+1:n}}, \tag{14}$$

$$\mu_k(\tau) = \frac{\det(J - \tau I)_{k:n}}{\det(J - \tau I)_{k+1:n}} - b_{k-1}c_{k-1}\frac{\det(J - \tau I)_{1:k-2}}{\det(J - \tau I)_{1:k-1}}, \tag{15}$$

$$\mu_k(\tau) = \frac{\det(J - \tau I)_{1:k}}{\det(J - \tau I)_{1:k-1}} + \frac{\det(J - \tau I)_{k:n}}{\det(J - \tau I)_{k+1:n}} - (a_k - \tau). \tag{16}$$

Remark. The above formulae are independent of any factorizations, and hence are superior to any other derived formulae which require the existence of factorizations of the submatrices. Also note that μ_k can approach a finite limit even if the minors in the denominator vanish provided the ratios of minors exist as limits.

COROLLARY 4.3.

$$\mu_k(\tau) = (a_k - \tau) - \frac{b_{k-1}c_{k-1}}{d_{k-1}} - \frac{b_kc_k}{\delta_{k+1}}, \tag{17}$$

$$\mu_k(\tau) = d_k - \frac{b_kc_k}{\delta_{k+1}}, \tag{18}$$

$$\mu_k(\tau) = \delta_k - \frac{b_{k-1}c_{k-1}}{d_{k-1}}, \tag{19}$$

$$\mu_k(\tau) = d_k + \delta_k - (a_k - \tau). \tag{20}$$

Proof. This is a direct consequence of Lemma 2.2. ∎

The way to derive the formulae for μ_k with nearest neighbor pivoting is obvious from Lemma 2.3.

Further formulae can be derived by replacing d_k by r-values of the QR factorizations and δ_k by l-values of the QL factorizations using Lemma 2.2.

COROLLARY 4.4.

$$\mu_k(\tau) = (a_k - \tau) - b_{k-1}\cos\alpha_{k-2}\tan\alpha_{k-1} - c_k\cos\beta_{k+2}\tan\beta_{k+1},$$

$$\mu_k(\tau) = \frac{\hat{r}_k}{\cos\alpha_{k-1}} - c_k\cos\beta_{k+2}\tan\beta_{k+1},$$

$$\mu_k(\tau) = \frac{\hat{l}_k}{\cos\beta_{k+1}} - b_{k-1}\cos\alpha_{k-2}\tan\alpha_{k-1},$$

$$\mu_k(\tau) = \frac{\hat{r}_k}{\cos\alpha_{k-1}} + \frac{\hat{l}_k}{\cos\beta_{k+1}} - (a_k - \tau).$$

The next result is stated for a more general matrix. Note that a tridiagonal matrix is both upper and lower Hessenberg.

THEOREM 4.5. *Let H be a Hessenberg matrix, and*

$$(H - \tau I)\boldsymbol{z} = \mu_k(\tau)\boldsymbol{e}_k, \quad z_k = 1.$$

The determinant of the matrix $H - \tau I$ is then given by

$$\det(H - \tau I) = \mu_k(\tau)\det(H - \tau I)_{1:k-1}\det(H - \tau I)_{k+1:n}.$$

Furthermore, if the LDU factorization of $(H - \tau I)_{1:k-1}$ and the UDL factorization of $(H - \tau I)_{k+1:n}$ exist, then the determinant of $H - \tau I$ is given by the determinant of the diagonal matrix

$$\operatorname{diag}(D_{1:k-1}, \mu_k, \Delta_{k+1:n}),$$

where $D_{1:k-1}$ and $\Delta_{k+1:n}$ denote the pivots of LDU and UDL factorizations of $(H - \tau I)_{1:k-1}$ and $(H - \tau I)_{k+1:n}$, respectively.

Proof. Cramer's rule gives

$$z_k = \mu_k(\tau)\frac{\det(H - \tau I)_{[k,k]}}{\det(H - \tau I)}.$$

Since $z_k = 1$ and $\det(H - \tau I)_{[k,k]} = \det(H - \tau I)_{1:k-1}\det(H - \tau I)_{k+1:n}$, we get the proposed result. ∎

The following result is immediate.

COROLLARY 4.6. *The eigenvalues of H are given by the zeros of $\mu_k(\tau)$,*

$$\mu_k(\tau) = \frac{\det\,(H - \tau I)}{\det\,(H - \tau I)_{1:k-1}\det\,(H - \tau I)_{k+1:n}}.$$

COROLLARY 4.7. $\mu_k\delta_{k+1} = d_k\mu_{k+1}$, $1 \le k \le n-1$.

Proof. From Theorem 4.5,

$$\det(H - \tau I) = \text{diag}\,(D_{1:k-1}, \mu_k, \Delta_{k+1:n}) = \text{diag}\,(D_{1:k}, \mu_{k+1}, \Delta_{k+2:n}).$$

A simple comparison of the two formulae for the determinant gives the stated result since d_i, $i = 1, \ldots, k-1$ and δ_i, $i = k+2, \ldots, n$ are not zero. ∎

The above corollary gives two more ways to compute μ_k.

$$\mu_{k+1} = \mu_k\delta_{k+1}/d_k, \quad k = 1, \ldots, n-1 \text{ with } \mu_1 = \delta_1, \tag{21}$$

$$\mu_k = \mu_{k+1}d_k/\delta_{k+1}, \quad k = n-1, \ldots, 1 \text{ with } \mu_n = d_n, \tag{22}$$

and these formulae are valid for Hessenberg matrices too. We have assumed the existence of the *LDU* and *UDL* factorizations of the matrix $H - \tau I$ to avoid division by zero in (21) and (22). The main advantage of the above formulae compared with formulae (17) to (20) is the absence of subtractions (hence no cancellations).

Example 1. We have computed[2] the μ_i-values for the Wilkinson matrix W_{21}^- using formula (19). See Table 1. For the zero shift, μ_i is zero for $i = 10, \ldots, 14$ and hence the computed determinant is zero according to Theorem 4.5. For the shift computed by the LAPACK routine SSTEQR (see [1]), no μ_i is zero. However, μ_{10} and μ_{12} are near zero to working precision (i.e., smaller than $\|J - \tau I\| * \text{macheps}$), and hence they could be thresholded to zero. Thus, the double factorization could be a more reliable method to determine near singularity of matrices. We have also computed the μ_i-values using (21). Note the disappearance of the zero values of μ_i when computed using this formula. We recall that according to Corollary 3.4, the μ_i cannot have zero values for a subset of indices i. Thus equation (21) could be used to replace zero values with more realistic estimates. However, note the change of sign of μ_i from $i = 9$ onward for the zero shift in the relevant columns in the table. This shows that formulae (21) and (22) should be used with extra care.

In floating-point arithmetic, the inertia counts determined by *LDU* and *UDL* factorizations of $J - \tau I$ might not be identical since the floating-point errors in computing the d_i and the δ_i will not be identical except for trivial problems. Thus, inconsistent results can be obtained if these formulae for the μ_k are used in an haphazard way. The erratic sign reversals can be avoided by using formula (21) for $i = 1, \ldots, k$ and formula (22) for $i = n, \ldots, k-1$ to compute μ_k with $k \approx 12$. Accuracy issues concerning the computation of the μ_k are studied in [6]. ∎

5. Computation of eigenvectors. The following result indicates how to compute good approximations to eigenvectors.

THEOREM 5.1. *If the LDU factorization of* $(J - \tau I)_{1:k-1}$ *exists with* $d_{k-1} \neq 0$ *and the UDL factorization of* $(J - \tau I)_{k+1:n}$ *exists with* $\delta_{k+1} \neq 0$, *and*

$$(J - \tau I)\boldsymbol{z} = \mu_k \boldsymbol{e}_k, \quad z_k = 1, \tag{23}$$

then

$$z_j = -(b_j/d_j)z_{j+1} \text{ for } j = k-1, \ldots, 1, \tag{24}$$

[2] We have used IEEE single precision arithmetic (24 bit mantissa) on an SGI workstation.

TABLE 1
μ_i-values for 11th eigenvalue of W_{21}^-.

τ	0	2.25453E−07	0	2.25453E−07
Eqn.	(19)	(19)	(21)	(21)
i	μ_i	μ_i	μ_i	μ_i
1	9.87915E+00	9.88732E+00	9.87915E+00	9.88732E+00
2	8.17441E+00	8.77475E+00	8.17441E+00	8.77475E+00
3	1.26584E+00	7.87189E+00	1.26584E+00	7.87189E+00
4	2.42356E−02	6.33561E+01	2.42356E−02	6.33561E+01
5	5.14835E−04	−1.63197E−01	5.14834E−04	−1.63197E−01
6	1.50204E−05	−4.63223E−03	1.50219E−05	−4.63223E−03
7	6.25849E−07	−1.98454E−04	6.44136E−07	−1.98438E−04
8	5.96046E−08	−1.37985E−05	4.47743E−08	−1.37928E−05
9	−2.98023E−08	−1.84774E−06	5.97978E−09	−1.84208E−06
10	0.00000E+00	−7.15256E−07	**2.23815E−09**	−6.89463E−07
11	0.00000E+00	−4.76837E−06	**1.48509E−08**	−4.57484E−06
12	0.00000E+00	−7.15256E−07	**2.23815E−09**	−6.89463E−07
13	0.00000E+00	−2.02656E−06	**5.97978E−09**	−1.84207E−06
14	0.00000E+00	−1.52588E−05	**4.47743E−08**	−1.37927E−05
15	−7.15256E−07	−2.20299E−04	6.44136E−07	−1.98414E−04
16	−1.47820E−05	−5.13029E−03	1.50219E−05	−4.62231E−03
17	−5.14984E−04	−1.70722E−01	5.14834E−04	−1.53811E−01
18	−2.42357E−02	−3.71081E+00	2.42356E−02	−3.34322E+00
19	−1.26584E+00	−7.62706E+00	1.26584E+00	−6.87153E+00
20	−8.17441E+00	−8.77079E+00	8.17441E+00	−7.90197E+00
21	−9.87915E+00	−9.88727E+00	9.87914E+00	−8.90785E+00

$$z_j = -(c_{j-1}/\delta_j)z_{j-1} \quad \textit{for} \quad j = k+1, \ldots, n. \tag{25}$$

Proof. Consider the LDU factorization $(J - \tau I)_{1:k-1} = L_{1:k-1}D_{1:k-1}U_{1:k-1}$. The first $k-1$ equations of (23) can be written in the form

$$L_{1:k-1}D_{1:k-1}U_{1:k-1}\boldsymbol{z}_{1:k-1} = -b_{k-1}z_k\boldsymbol{e}_{k-1} = -b_{k-1}\boldsymbol{e}_{k-1}.$$

Thus,

$$D_{1:k-1}U_{1:k-1}\boldsymbol{z}_{1:k-1} = -b_{k-1}\boldsymbol{e}_{k-1},$$

$$U_{1:k-1}\boldsymbol{z}_{1:k-1} = -(b_{k-1}/d_{k-1})\boldsymbol{e}_{k-1},$$

which gives (24). Similarly, by considering the trailing $n-k$ equations of (23) we get (25). □

Remark 1. Once k is chosen, it is not necessary to know the value of μ_k for the solution of the system of equations (23). In fact, it is possible to compute μ_k once $z_{k-1} = -b_{k-1}/d_{k-1}$ and $z_{k+1} = -c_k/\delta_{k+1}$ are known. See (11) and (17).

Remark 2. Since μ_k is not required for the solution of (23) and we are not interested in the norm of the solution $\boldsymbol{y}$, instead of (23) one may solve the set of equations

$$(J - \tau I)\boldsymbol{y} = \alpha\boldsymbol{e}_k$$

for any convenient nonzero constant α (e.g., α set to a small value to avoid overflow problems). If the constant term is chosen as unity, then $\boldsymbol{y}$ is given by the kth column

TABLE 2
Computed eigenvectors in single precision.

τ	1.07461942E+01			
Eqn.	(19)	(24),(25)	(24),(25)	
Drop k		1	21	
i	μ_i	z_i	$\hat{z}_i$	Difference
1	4.768372E−07	1.000000E+00	1.000000E+00	0.000000E+00
2	8.344650E−07	7.461944E−01	7.461939E−01	4.768372E−07
3	5.245209E−06	3.030000E−01	3.029992E−01	8.046627E−07
4	6.556511E−05	8.590253E−02	8.590069E−02	1.832843E−06
5	1.370430E−03	1.880749E−02	1.880145E−02	6.042421E−06
6	4.323769E−02	3.361466E−03	3.334617E−03	2.684933E−05
7	2.650068E+00	5.081474E−04	3.599074E−04	1.482400E−04
8	−8.027469E+00	6.659435E−05	−9.066119E−04	9.732062E−04
9	−8.519791E+00	7.705368E−06	−7.382698E−03	7.390403E−03
10	−9.536426E+00	7.982861E−07	−6.366390E−02	6.366470E−02
11	−1.055664E+01	7.488273E−08	−6.130980E−01	6.130981E−01
12	−1.157332E+01	6.418178E−09	−6.524806E+00	6.524806E+00
13	−1.258726E+01	5.064417E−10	−7.602853E+01	7.602853E+01
14	−1.359910E+01	3.702578E−11	−9.625497E+02	9.625497E+02
15	−1.460928E+01	2.521772E−12	−1.315537E+04	1.315537E+04
16	−1.561812E+01	1.607630E−13	−1.930290E+05	1.930290E+05
17	−1.662589E+01	9.632481E−15	−3.026317E+06	3.026317E+06
18	−1.763276E+01	5.444324E−16	−5.048626E+07	5.048626E+07
19	−1.863889E+01	2.912115E−17	−8.929124E+08	8.929124E+08
20	−1.964449E+01	1.478382E−18	−1.668822E+10	1.668822E+10
21	−2.069541E+01	7.126039E−20	−3.286360E+11	3.286360E+11

of $(J - \tau I)^{-1}$, which is a step of inverse iteration with the right-hand side of the equation set to $\boldsymbol{e}_k$.

Remark 3. If $d_{i-1} = 0$, then $d_i = \pm\infty$ for a particular i, $i \leq k-1$. In that case, $z_i = -(b_i/d_i)z_{i+1} = 0$. The element z_{i-1} can be obtained via $z_{i-1} = -(b_{i-1}/d_{i-1})z_i = ((b_{i-1}b_i)/(d_{i-1}d_i))z_{i+1}$ and noting that $d_{i-1}d_i = -b_{i-1}c_{i-1}$ in the limit. See Lemma 8.2. Then $z_{i-1} = -(b_i/c_{i-1})z_{i+1}$. Similar formulae can be derived if $\delta_{i+1} = 0$ for a particular i, $i \geq k+1$.

Remark 4. In floating-point arithmetic, the d_i and similarly the δ_i should be thresholded to θ if these quantities are tiny

$$\text{if } \theta \leq d_i \leq -\theta \text{ then } d_i \leftarrow \theta, \quad \theta \geq \eta,$$

where η is the smallest representable number in the machine. See Kahan [14] for further details.

If $|\mu_k|$ is tiny, we could expect very good approximations to eigenvectors from Theorem 5.1. However, if $|\mu_k|$ is not tiny then we are computing the eigenvectors of the perturbed matrix $J - \mu_k(\tau)\boldsymbol{e}_k\boldsymbol{e}_k^*$, and in that case the computed eigenvectors will not closely approximate the eigenvectors of J.

Example 2. We have repeated an experiment done by Wilkinson [17], [18], [19] for the matrix W_{21}^-. Table 2 shows the computed eigenvectors corresponding to the largest eigenvalue in IEEE single precision arithmetic. The vectors $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$ denote the eigenvectors computed by dropping the first and the 21st equations, respectively. The vector $\hat{\boldsymbol{z}}$ is scaled such that the first element is unity. Since $|\mu_1|$ is the smallest and $|\mu_{21}|$ is the largest, $\boldsymbol{z}$ and the $\hat{\boldsymbol{z}}$ represent the best and the worst possible solutions. Note that the worst case does not have any correspondence to the best case.

Table 3 shows the same results in IEEE double precision arithmetic (53-bit man-

TABLE 3
Computed eigenvectors in double precision.

τ	1.0746194182903357D+01			
Eqn.	(19)	(24),(25)	(24),(25)	
Drop k		1	21	
i	μ_i	z_i	$\hat{z}_i$	Difference
1	3.330669D−16	1.000000D+00	1.000000D+00	0.000000D+00
2	6.661338D−16	7.461942D−01	7.461942D−01	2.220446D−16
3	3.552714D−15	3.029999D−01	3.029999D−01	5.551115D−16
4	4.218847D−14	8.590249D−02	8.590249D−02	1.179612D−15
5	8.766321D−13	1.880748D−02	1.880748D−02	3.864964D−15
6	2.743761D−11	3.361465D−03	3.361465D−03	1.717550D−14
7	1.200664D−09	5.081471D−04	5.081471D−04	9.482627D−14
8	6.990788D−08	6.659431D−05	6.659431D−05	6.225412D−13
9	5.221731D−06	7.705362D−06	7.705358D−06	4.727499D−12
10	4.865260D−04	7.982854D−07	7.982447D−07	4.072508D−11
11	5.557997D−02	7.488266D−08	7.449047D−08	3.921870D−10
12	2.152249D+01	6.418173D−09	2.244380D−09	4.173793D−09
13	−1.271972D+01	5.064412D−10	−4.812755D−08	4.863399D−08
14	−1.359992D+01	3.702574D−11	−6.156875D−07	6.157245D−07
15	−1.460928D+01	2.521769D−12	−8.415233D−06	8.415235D−06
16	−1.561812D+01	1.607628D−13	−1.234770D−04	1.234770D−04
17	−1.662589D+01	9.632470D−15	−1.935877D−03	1.935877D−03
18	−1.763276D+01	5.444317D−16	−3.229510D−02	3.229510D−02
19	−1.863889D+01	2.912112D−17	−5.711792D−01	5.711792D−01
20	−1.964449D+01	1.478380D−18	−1.067514D+01	1.067514D+01
21	−2.069541D+01	7.126029D−20	−2.102222D+02	2.102222D+02

tissa). Because of the improved accuracy of the shift and perhaps also due to the higher precision of the arithmetic, the smallest $|\mu_i|$ are considerably smaller than in single precision. Also note that the difference between the best and the worst case is converging as the precision goes up. By comparing $\boldsymbol{z}$ computed in single precision and in double precision, it can be seen that the single precision result is very accurate.

The 2-norm of the matrix is equal to the 21st eigenvalue. In both precisions, $|\mu_1|$ is smaller than $\|J\|_2 * \text{macheps}$. Thus $\boldsymbol{z}$ should be a good approximation to the 21st eigenvector. □

Example 3. We have contrived a tridiagonal matrix to illustrate the form of μ_i-values for nonsymmetric matrices. In the Wilkinson matrix W_{21}^-, all off-diagonal elements are equal to unity. We define a new matrix by setting

$$c_i = 1, \;\; b_i = -1 \;\text{ for }\; i = 11, \ldots, 20,$$

$$c_i = -1, \;\; b_i = 1 \;\text{ for }\; i = 1, \ldots, 10$$

but with the same diagonal values as the W_{21}^- matrix. This new tridiagonal, which we call the unsymmetric W matrix, has two pairs of complex conjugate eigenvalues. We have computed in IEEE single precision the μ_i-values for one of the complex eigenvalues (shifts) τ. See Table 4. The minimal μ_k is when $k = 20$. The optimal k for the eigenvalue $\bar{\tau}$ is the same as for τ; see Corollary 3.6. □

Although Theorem 5.1 gives the principal algorithm for computation of eigenvectors, there are other secondary ways to compute them. Suppose that z_{k-1} was computed using (24), then it is possible to compute the rest of the elements z_j, $j = k-2, \ldots, 1$ via the three-term recurrence

$$c_{j-2}z_{j-2} + (a_{j-1} - \tau)z_{j-1} + b_{j-1}z_j = 0.$$

TABLE 4
Computed μ_i-values for the unsymmetric W matrix.

τ	$-9.056515\ -\imath\ 0.7829880$		
i	real $\{\mu_i\}$	imag $\{\mu_i\}$	$\lvert\mu_i\rvert$
1	1.911162E+01	7.806144E−01	1.912755E+01
2	1.816720E+01	7.781793E−01	1.818386E+01
3	1.717351E+01	7.776191E−01	1.719111E+01
4	1.618077E+01	7.769318E−01	1.619941E+01
5	1.518897E+01	7.761040E−01	1.520879E+01
6	1.419833E+01	7.750947E−01	1.421947E+01
7	1.320911E+01	7.738469E−01	1.323175E+01
8	1.222165E+01	7.722794E−01	1.224602E+01
9	1.123641E+01	7.702739E−01	1.126278E+01
10	1.025406E+01	7.676529E−01	1.028275E+01
11	9.275450E+00	7.641196E−01	9.306871E+00
12	8.304144E+00	7.610373E−01	8.338943E+00
13	7.246818E+00	6.419530E−01	7.275196E+00
14	3.339067E+00	6.002015E+00	6.868301E+00
15	7.866935E−02	−1.669632E−01	1.845686E−01
16	−7.753968E−04	6.662467E−03	6.707437E−03
17	−8.402765E−05	−3.510378E−04	3.609546E−04
18	1.999736E−05	2.475828E−05	3.182557E−05
19	−5.066395E−06	−1.542270E−06	5.295938E−06
20	1.728535E−06	−1.102686E−06	2.050304E−06
21	1.072884E−06	2.920628E−06	3.111454E−06

See the proof of Lemma 3.2 for details. Similarly, if z_{k+1} is determined by (25), the above recursion could be used to compute z_j for $j = k+2, \ldots, n$.

6. Perturbation of the offdiagonals. We have already considered the case of perturbing a diagonal element of J to make the perturbed matrix singular. However, for tridiagonal matrices with zero diagonals, such perturbations will destroy the matrix structure. It is well known that zero diagonal tridiagonal matrices are paramount in the study of the SVD of bidiagonal matrices. See Golub and Kahan [9], Kahan [14], Demmel and Kahan [5]. They are also important in vibration analysis. See Bishop et al. [3].

It is possible to perturb the product $b_k c_k$ for a particular k such that $\hat{\mu}_k$ (the μ_k of the perturbed matrix) is zero in which case the perturbed matrix is singular. Alternatively, the product $b_{k-1}c_{k-1}$ can be perturbed such that $\tilde{\mu}_k$ (the μ_k of the perturbed matrix) is zero.

THEOREM 6.1. *Consider the perturbation of the off-diagonal element product of J,*

$$b_k c_k \to \hat{b}_k \hat{c}_k = b_k c_k + \hat{\rho}_k(\tau)$$

for a particular k, $1 \le k \le n-1$. Then $\hat{\mu}_k(\tau)$ (the $\mu_k(\tau)$ of the perturbed matrix) is zero if

$$\hat{\rho}_k(\tau) = \mu_k(\tau)\delta_{k+1}. \tag{26}$$

Similarly, if the product $b_{k-1}c_{k-1}$ is perturbed,

$$b_{k-1}c_{k-1} \to \tilde{b}_{k-1}\tilde{c}_{k-1} = b_{k-1}c_{k-1} + \tilde{\rho}_{k-1}(\tau), \quad 2 \le k \le n,$$

then $\tilde{\mu}_k(\tau)$ (the $\mu_k(\tau)$ of the perturbed matrix) is zero if

$$\tilde{\rho}_{k-1}(\tau) = \mu_k d_{k-1}(\tau). \tag{27}$$

Furthermore,

$$\tilde{\rho}_k(\tau) = \hat{\rho}_k(\tau), \;\; 1 \le k \le n-1.$$

Proof. For the unperturbed matrix, equation (18) is

$$\mu_k = d_k - b_k c_k/\delta_{k+1}$$

and for the perturbed case

$$\hat{\mu}_k = 0 = d_k - (b_k c_k + \hat{\rho}_k)/\delta_{k+1}.$$

A simple comparison of the above two equations gives the first result

$$\hat{\rho}_k = \mu_k \delta_{k+1}, \;\; 1 \le k \le n-1. \tag{28}$$

Similarly, equation (19) leads to the second result

$$\tilde{\rho}_{k-1} = \mu_k d_{k-1}, \;\; 2 \le k \le n,$$

which could be also written in the form

$$\tilde{\rho}_k = \mu_{k+1} d_k, \;\; 1 \le k \le n-1.$$

Using Corollary 4.7

$$\tilde{\rho}_k = \mu_{k+1} d_k = \mu_k \delta_{k+1}, \;\; 1 \le k \le n-1.$$

By comparing the above equation with (28) we arrive at the final result. ∎

Remark 1. Since $\hat{\rho}_k(\tau) = \tilde{\rho}_k(\tau)$, it is no longer necessary to have embellishments over the ρ_k. However, in inexact arithmetic different formulae for ρ_k could give different results.

Remark 2. Recall that

$$\hat{b}_k \hat{c}_k = b_k c_k + \rho_k$$

and hence

$$(\hat{b}_k/b_k)(\hat{c}_k/c_k) = 1 + \rho_k/(b_k c_k).$$

If the relative perturbations of b_k and c_k are equal, $(\hat{b}_k/b_k) = (\hat{c}_k/c_k)$, then

$$(\hat{b}_k/b_k) = (\hat{c}_k/c_k) = \sqrt{1 + \rho_k/(b_k c_k)}.$$

7. Quality of the computed eigenvalues and eigenvectors. The proofs of the first two theorems in this section can be found in [7] for any square matrix.

THEOREM 7.1. *Let J be a matrix with linear elementary divisors*

$$(J - \tau I)\boldsymbol{z}(k,\tau) = \mu_k(\tau)\boldsymbol{e}_k, \;\; z_k = 1,$$

where τ is not an eigenvalue of J. Then

$$\min_i |\lambda_i - \tau| \le |\mu_k(\tau)|||X||_2 ||\boldsymbol{y}_k||_2/||\boldsymbol{z}||_2,$$

where $J = X\Lambda Y$, $Y = X^{-1}$, *and* $\boldsymbol{x}^{(i)}$ *denotes the* i*th column of* X. *Furthermore, if* J *is normal, then*

$$\min_i |\lambda_i - \tau| \leq |\mu_k(\tau)|/||\boldsymbol{z}||_2.$$

COROLLARY 7.2. *If* $\mu_k(\tau) \to 0$, *then* $\min_i |\lambda_i - \tau| \to 0$.

The angle between the computed eigenvector $\boldsymbol{z}$ and the desired eigenvector $\boldsymbol{x}$ is a good indicator of the quality of the approximation.

THEOREM 7.3. *Let* J *be a normal matrix,* $(J-\tau I)\boldsymbol{z} = \mu_k(\tau)\boldsymbol{e}_k$, *and* $z_k = 1$ *where* τ *is not an eigenvalue of* J. *Then the angle between* $\boldsymbol{z}$ *and the eigenvector* $\boldsymbol{x}(= \boldsymbol{x}^{(j)})$ *corresponding to the eigenvalue* $\lambda(= \lambda_j)$ *nearest to* τ *is given by*

$$|\cos\angle| = \frac{|\boldsymbol{x}^*\boldsymbol{z}|}{||\boldsymbol{x}||_2||\boldsymbol{z}||_2} = 1\Big/\left(1 + \sum_{i\neq j}\left|\frac{x_k^{(i)}}{x_k^{(j)}}\frac{\lambda-\tau}{\lambda_i-\tau}\right|^2\right)^{\frac{1}{2}}$$

if $x_k^{(j)} \neq 0$.

Remark 1. This theorem confirms the Wilkinson's analysis [17], which specifies that k should be chosen such that $|x_k|$ is maximal if a good approximation to an eigenvector is required.

Remark 2. This theorem also shows that if the gap $= \min_{i\neq j}|\lambda - \lambda_i|$ is large, then the computed eigenvector $\boldsymbol{z}$ is a good approximation to $\boldsymbol{x}$.

Another way to compare the accuracy of the computed eigenvector $\boldsymbol{z}$ is to estimate $||\boldsymbol{z}-\boldsymbol{x}||_{1:k-1}$ and $||\boldsymbol{z}-\boldsymbol{x}||_{k+1:n}$, where $\boldsymbol{x}$ is the desired eigenvector with $x_k = 1$.

THEOREM 7.4. *Let* $(J-\tau I)\boldsymbol{z} = \mu_k(\tau)\boldsymbol{e}_k$, $z_k = 1$ *where* τ *is not an eigenvalue of* J. *Then*

$$\frac{|\lambda-\tau|\,||\boldsymbol{x}_{1:k-1}||}{||(J-\tau I)_{1:k-1}||} \leq ||(\boldsymbol{z}-\boldsymbol{x})_{1:k-1}|| \leq |\lambda-\tau|\,||\{(J-\tau I)_{1:k-1}\}^{-1}||\,||\boldsymbol{x}_{1:k-1}||,$$

$$\frac{|\lambda-\tau|\,||\boldsymbol{x}_{k+1:n}||}{||(J-\tau I)_{k+1:n}||} \leq ||(\boldsymbol{z}-\boldsymbol{x})_{k+1:n}|| \leq |\lambda-\tau|\,||\{(J-\tau I)_{k+1:n}\}^{-1}||\,||\boldsymbol{x}_{k+1:n}||,$$

where $\boldsymbol{x}$ *is an eigenvector of* J *with* $x_k = 1$ *and* λ *is the corresponding eigenvalue.*

Proof. From $(J-\tau I)\boldsymbol{z} = \mu_k(\tau)\boldsymbol{e}_k$, we get

$$(J-\tau I)(\boldsymbol{z}-\boldsymbol{x}) = (\tau-\lambda)\boldsymbol{x} + \mu_k(\tau)\boldsymbol{e}_k. \tag{29}$$

By removing the kth equation of (29), we obtain

$$(J-\tau I)_{[k,k]}(\boldsymbol{z}-\boldsymbol{x})_{[k]} = (\tau-\lambda)\boldsymbol{x}_{[k]}. \tag{30}$$

However, the submatrices $J_{1:k-1,k+1:n}$ and $J_{k+1:n,1:k-1}$ are null matrices and hence we get two decoupled systems

$$(J-\tau I)_{1:k-1}(\boldsymbol{z}-\boldsymbol{x})_{1:k-1} = (\tau-\lambda)\boldsymbol{x}_{1:k-1}, \tag{31}$$

$$(J-\tau I)_{k+1:n}(\boldsymbol{z}-\boldsymbol{x})_{k+1:n} = (\tau-\lambda)\boldsymbol{x}_{k+1:n}. \tag{32}$$

By taking the norms of the last two equations we get the lower bounds. If the inverse of the matrix $(J-\tau I)_{1:k-1}$ exists then from (31),

$$(\boldsymbol{z}-\boldsymbol{x})_{1:k-1} = (\tau-\lambda)\{(J-\tau I)_{1:k-1}\}^{-1}\boldsymbol{x}_{1:k-1},$$

which gives one of the upper bounds. Similarly, the second upper bound can be obtained using (32). ∎

An immediate corollary is as follows.

COROLLARY 7.5.

$$\frac{|\lambda-\tau|\ ||\boldsymbol{x}_{1:k-1}||_2}{\sigma_{\max}\{(J-\tau I)_{1:k-1}\}} \le ||(\boldsymbol{z}-\boldsymbol{x})_{1:k-1}||_2 \le \frac{|\lambda-\tau|\ ||\boldsymbol{x}_{1:k-1}||_2}{\sigma_{\min}\{(J-\tau I)_{1:k-1}\}},$$

$$\frac{|\lambda-\tau|\ ||\boldsymbol{x}_{k+1:n}||_2}{\sigma_{\max}\{(J-\tau I)_{k+1:n}\}} \le ||(\boldsymbol{z}-\boldsymbol{x})_{k+1:n}||_2 \le \frac{|\lambda-\tau|\ ||\boldsymbol{x}_{k+1:n}||_2}{\sigma_{\min}\{(J-\tau I)_{k+1:n}\}}.$$

Remark. This corollary indicates that good approximations can be obtained if the minimum singular values of $(J-\tau I)_{1:k-1}$ and $(J-\tau I)_{k+1:n}$ are large.

8. Deflation using an ideal shift. It is well known that it is possible to deflate a matrix using a shift exactly equal to an eigenvalue of that matrix provided that exact arithmetic is used. The following result indicates how this could be achieved using the *LDU* factorization.

LEMMA 8.1. *If the LDU factorization of the matrix $J-\tau I$ exists where the shift τ is an exact eigenvalue of J, then d_n is zero.*

Proof. If the factorization exists, then none of the d_i (except perhaps d_n) can be zero. However, the determinant of $J-\tau I$ is given by the product of the d_i which is zero for an exact eigenvalue. Hence the determinant can vanish only if d_n is zero. ∎

We now show that the above result is true in the limit even if the factorization does not exist.

LEMMA 8.2. *Suppose that the LDU factorization of $J-\tau I$ does not exist as $\tau \to \hat{\tau}$ since $d_{i-1} \to 0$ for a particular i, $2 \le i \le n$ where $\hat{\tau}$ is an exact eigenvalue of J. Then, in the limiting case, $d_n \to 0$ as $\tau \to \hat{\tau}$.*

Proof. By multiplying (3) by d_{i-1},

$$d_i d_{i-1} = (a_i-\tau)d_{i-1} - b_{i-1}c_{i-1}$$

and by taking the limit

$$d_i d_{i-1} \to -b_{i-1}c_{i-1} \ \text{ as } \ \tau \to \hat{\tau}.$$

Thus the determinant is finite in the limit, and by using an argument similar to Lemma 8.1 we get the stated result. ∎

Remark. If there is more than one i such that $d_{i-1} \to 0$, then the limit can be evaluated repeatedly for each i.

In floating-point arithmetic, the assertion that the determinant is zero can break down due to two reasons. First, in general, an eigenvalue is not known or even representable to full accuracy. Second, because of rounding errors the ideal outcome, $d_n = 0$, can happen only accidentally. However, an optimist might expect that if the shift represents an eigenvalue to its full machine precision, then d_n and hence the determinant would be tiny.

TABLE 5
Pivots for the 11th eigenvalue of W_{21}^-.

τ	0	2.25453E−07	
$\nu(\tau)$	11	11	
eqn.	(3)	(3)	
i	d_i	d_i	difference
21	−9.87915E+00	−9.88727E+00	8.12531E−03

TABLE 6
Pivots for the 13th eigenvalue of W_{21}^-.

τ	1.99999988E+00	2.0000000E+00	
$\nu(\tau)$	12	13	
eqn.	(3)	(3)	
i	d_i	d_i	difference
21	−1.19082E+01	−1.19082E+01	0.00000E+00

If d_n is zero (or tiny), then the following result provides an algorithm to compute an eigenvector $\boldsymbol{x}$ of J. The proof is straightforward.

THEOREM 8.3. *If τ is an exact eigenvalue of J, then the eigenvector corresponding to the eigenvalue τ is given by*

$$x_j = -(b_j/d_j)x_{j+1} \text{ for } j = n-1, \ldots, 1 \text{ with } x_n = 1.$$

Remark. The residual vector corresponding to the vector x is given by $d_n \boldsymbol{e}_n$ and hence the norm of this residual is $|d_n|$. To obtain a zero residual it is necessary and sufficient that d_n is zero.

Example 4. We have computed the recurrence for the Wilkinson matrix W_{21}^- with the 11th eigenvalue as the shift τ. This eigenvalue is zero in exact arithmetic and hence known to full precision. The second column of Table 5 gives the d_n with the shift equal to zero, and the third column gives the d_n with the shift set to the eigenvalue given by the LAPACK routine SSTEQR. See [1] for a description of this routine. It can be seen that d_n is not zero or tiny in either column. □

Example 5. Table 6 shows the pivot d_n for the Wilkinson W_{21}^- matrix for two values of the shift which are separated by one *ulp* (units in the last place held). These two shifts straddle the 13th eigenvalue, which can be observed by the change in the inertia $\nu(\tau)$ which counts the negative d_i. Again, the d_n do not vanish as prescribed by Lemma 8.1. □

If the determinant is zero, then it indicates a singular matrix. However, nearly singular matrices (i.e., the smallest singular value is tiny) do not always have nearly zero determinants. See section 2.7.3 of Golub and Van Loan [10] for further details. In Examples 3 and 4, the determinants (as computed by $d_1 d_2 \ldots d_n$) are not zero (in fact they are huge) although the matrices are almost singular.

Theorem 8.3 gives an algorithm for computing eigenvectors which relies on the fact that d_n is zero. In Examples 3 and 4, the computed d_n is far from zero. There are many unanswered questions. Is it reasonable to expect d_n to be zero or tiny in floating-point arithmetic? Is it due to "forward instability" that d_n is not zero?

We have proved that with an exact eigenvalue shift, d_n and the determinant are zero in exact arithmetic. How do we measure this deviation from the mathematically ideal case when the shift is not an exact eigenvalue and when the arithmetic is not

exact?

Note that the forward recurrence (3) can be written as the backward recurrence

$$d_{i-1} = b_{i-1}c_{i-1}/(a_i - \tau - d_i), \quad i = n, \ldots, 2. \tag{33}$$

In the ideal case with an exact eigenvalue shift, $d_n = 0$. Suppose that we assume this ideal initial value $d_n = 0$ and run the recurrence (33) backward. In exact arithmetic with the exact eigenvalue shift, both (3) and (33) should give identical d_i provided the *LDU* factorization exists. However, for a general shift τ, the d_i given by (33) will be different from that of (3). To differentiate these two d_i recurrences, the one which goes in the backward direction will be renamed $\check{d}_i$.

$$\check{d}_n = 0, \quad \check{d}_{i-1} = b_{i-1}c_{i-1}/(a_i - \tau - \check{d}_i), \quad i = n, \ldots, 2. \tag{34}$$

This backward recurrence (34) for the $\check{d}_i$ is probably unfamiliar to some readers although similar recurrences are not so uncommon in continued fractions literature. See Chapter 12 of Wall [16].

Our objective is to find out how the forward recurrence and the backward recurrence differ by comparing each d_i with the corresponding $\check{d}_i$.

DEFINITION 8.4. *The deviation of each d_i (given by (3)) from $\check{d}_i$ (as given by (34)) is defined as $\omega_i(\tau)$:*

$$\omega_i(\tau) = d_i - \check{d}_i, \quad i = 1, \ldots, n. \tag{35}$$

By definition, $\omega_n(\tau) = d_n = 0$ and hence $\omega_n(\tau)$ is a candidate in studying the problem of nonvanishing d_n.

THEOREM 8.5. *The deviation $\omega_i(\tau)$ is given by*

$$\omega_i(\tau) = d_i + \delta_i - (a_i - \tau), \quad i = 1, \ldots, n, \tag{36}$$

where the d_i are the pivots of the LDU factorization of $(J - \tau I)_{1:i}$ and the δ_i are the pivots of the UDL factorization of $(J - \tau I)_{i:n}$. Furthermore, $\omega_i(\tau) = \mu_i(\tau)$ where $\mu_i(\tau)$ is as defined in Theorem 3.3.

Proof. Consider the transformation

$$f_i = -\check{d}_i + a_i - \tau, \quad i = 1, \ldots, n. \tag{37}$$

The backward recurrence (34) can be written in terms of the f_i as

$$f_{i-1} = a_{i-1} - \tau - b_{i-1}c_{i-1}/f_i, \quad i = n, \ldots, 2. \tag{38}$$

The assumed initial condition $\check{d}_n = 0$ can be translated to f_n using (37) to give

$$\delta_n = a_n - \tau. \tag{39}$$

It is not difficult to recognize that the f_i recursion (38), together with the initial condition (39), give the diagonal pivots of the *UDL* factorization of $J - \tau I$. That is, $\delta_i = f_i$. See (4). Equation (36) then follows from (37) and (35). By comparing (36) with (20), we get the stated equality. ∎

The deviation $\omega_i(\tau)$ (or, equivalently, $\mu_i(\tau)$) for a particular i is not a totally new variable. In fact, it was defined by Babuška in his study of the numerical stability of tridiagonal solvers. However, this variable is well hidden in a set of 19 equations

pertaining to two-sided elimination. In his notation, $\omega_s(0)$ is π_s. See equation (5.29) of [2]. Since π is often used as the positive inertia count of a Hermitian matrix (i.e., the number of positive d_i or δ_i), we avoid his notation.

We have already established that, in exact arithmetic, $\mu_i(\tau) = 0$ for all i if τ is an exact eigenvalue of J. However, in inexact arithmetic and when the shift τ is not identical to an eigenvalue, all the $\mu_i(\tau)$ might not be zero or tiny.

We recall that the vanishing d_n was proved by considering the determinant of $J - \tau I$ with the eigenvalue shift τ. In particular, we used the formula

$$\det(J - \tau I) = d_1 \dots d_{n-1} d_n = d_1 \dots d_{n-1} \mu_n$$

to prove that d_n is zero if the *LDU* factorization of $J - \tau I$ exists. Similarly, it is possible to show that δ_1 is zero if the *UDL* factorization exists since

$$\det(J - \tau I) = \delta_1 \delta_2 \dots \delta_n = \mu_1 \delta_2 \dots \delta_n.$$

However, Theorem 4.5 gives further $n - 2$ formulae for the computation of the determinant

$$\det(J - \tau I) = d_1 \dots d_{k-1} \mu_k \delta_{k+1} \dots \delta_n, \quad k = 1, \dots, n. \tag{40}$$

We have proved that μ_k is zero for any k if the shift τ is an exact eigenvalue of J. See Corollary 3.5. In inexact arithmetic, instead of expecting a tiny d_n or δ_1, it is more reasonable to watch for a tiny μ_k with k chosen such that $|\mu_k| = \min_i^n |\mu_i|$. See Example 1 and Table 1 where $\mu_k = 0$ for $k = 10, \dots, 14$. Thus, we get a zero determinant without a vanishing d_n in floating-point arithmetic.

Acknowledgments. The author wishes to thank Professor Gene Golub for the Babuška reference, Professor G.W. "Pete" Stewart for the Cauchy reference, and Professor Bill Gragg for helping to rediscover the Henrici reference. Many fruitful discussions with Jeremy Du Croz are gratefully acknowledged. His diligent comments led to many improvements. The final draft was read by Neil Swindells.

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, Release* 2.0, SIAM, Philadelphia, 1995.

[2] I. Babuška, *Numerical stability in problems of linear algebra*, SIAM J. Numer. Anal., 9 (1972), pp. 53–77.

[3] R. E. D. Bishop, G. M. L. Gladwell, and S. Michaelson, *Matrix Analysis of Vibrations*, Cambridge University Press, Cambridge, 1965.

[4] A. L. Cauchy, *Sur l'équation á l'aide de laquelle on détermine les inégalités séculaires des mouvements des planètes*, in Oeuvres Complétes (II^e Série), Vol. 9, Gauthier–Villars, Paris, 1841.

[5] J. Demmel and W. Kahan, *Accurate singular values of bidiagonal matrices*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 873–912.

[6] K. V. Fernando, *Accurate BABE Factorisation of Tridiagonal Matrices for Eigenproblems*, Technical Report TR5/95, Numerical Algorithms Group Ltd, Wilkinson House, Jordan Hill, Oxford, 1995.

[7] K. V. Fernando, *On a Classical Method for Computing Eigenvectors*, Technical Report TR3/95, Numerical Algorithms Group Ltd, Wilkinson House, Jordan Hill, Oxford, 1995.

[8] K. V. Fernando, *Computing an Eigenvector of a Tridiagonal When the Eigenvector is Known*, in Proc. of the Third International Congress on Industrial and Applied Mathematics, Zeitschrift für Angewandte Mathematik und Mechanik, Hamburg, 1996.

[9] G. H. Golub and W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, SIAM J. Numer. Anal., 2 (1965), pp. 205–224.
[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1989.
[11] P. Henrici, *Bounds for eigenvalues of certain tridiagonal matrices*, J. Soc. Indust. Appl. Math., 1 (1963), pp. 281–290.
[12] H. Holzer, *Die Berechnung der Drehschwingungen und ihre Anwendung im Maschinenbau*, Verlag von Julius Springer, Berlin, 1921.
[13] E. R. Jessup and I. C. F. Ipsen, *Improving the accuracy of inverse iteration*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 550–571.
[14] W. Kahan, *Accurate Eigenvalues of a Symmetric Tri-Diagonal Matrix*, Technical Report CS 41, Computer Science Department, Stanford University, Stanford, CA, 1966.
[15] J. Sherman and W. J. Morrison, *Adjustment of an inverse matrix corresponding to a change in one element of a given matrix*, Ann. Math. Statist., 21 (1950), pp. 124–126.
[16] H. S. Wall, *Analytic Theory of Continued Fractions*, Van Nostrand, New York, 1948.
[17] J. H. Wilkinson, *The calculation of eigenvectors of codiagonal matrices*, Computer J., 1 (1958), pp. 90–96.
[18] J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Her Majesty's Stationary Office, London, 1963.
[19] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 4, pp. 1035–1064, October 1997

015

ROBUST SOLUTIONS TO LEAST-SQUARES PROBLEMS WITH UNCERTAIN DATA *

LAURENT EL GHAOUI† AND HERVÉ LEBRET†

Abstract. We consider least-squares problems where the coefficient matrices A, b are unknown but bounded. We minimize the worst-case residual error using (convex) second-order cone programming, yielding an algorithm with complexity similar to one singular value decomposition of A. The method can be interpreted as a Tikhonov regularization procedure, with the advantage that it provides an exact bound on the robustness of solution and a rigorous way to compute the regularization parameter. When the perturbation has a known (e.g., Toeplitz) structure, the same problem can be solved in polynomial-time using semidefinite programming (SDP). We also consider the case when A, b are rational functions of an unknown-but-bounded perturbation vector. We show how to minimize (via SDP) upper bounds on the optimal worst-case residual. We provide numerical examples, including one from robust identification and one from robust interpolation.

Key words. least-squares problems, uncertainty, robustness, second-order cone programming, semidefinite programming, ill-conditioned problem, regularization, robust identification, robust interpolation

AMS subject classifications. 15A06, 65F10, 65F35, 65K10, 65Y20

PII. S0895479896298130

Notation. For a matrix X, $\|X\|$ denotes the largest singular value and $\|X\|_F$ the Frobenius norm. If x is a vector, $\max_i |x_i|$ is denoted by $\|x\|_\infty$. For a matrix A, $A^\dagger$ denotes the Moore–Penrose pseudoinverse of A. For a square matrix S, $S \geq 0$ (resp., $S > 0$) means S is symmetric and positive semidefinite (resp., definite). For $S \geq 0$, $S^{1/2}$ denotes the symmetric square root of S. For $S > 0$, and given vector x, we define $\|x\|_S = \|S^{-1/2}x\|$. The notation I_p denotes the $p \times p$ identity matrix; sometimes the subscript is omitted when it can be inferred from context. For given matrices X, Y, the notation $X \oplus Y$ refers to the block-diagonal matrix with X, Y as diagonal blocks.

1. Introduction.

Consider the problem of finding a solution x to an overdetermined set of equations $Ax \simeq b$, where the data matrices $A \in \mathbf{R}^{n\times m}$, $b \in \mathbf{R}^n$ are given. The least squares (LS) fit minimizes the residual $\|\Delta b\|$ subject to $Ax = b + \Delta b$, resulting in a consistent linear model of the form $(A, b + \Delta b)$ that is closest to the original one (in the Euclidean norm sense). The total least squares (TLS) solution described by Golub and Van Loan [17] finds the smallest error $\|[\Delta A\ \Delta b]\|_F$ subject to the consistency equation $(A + \Delta A)x = b + \Delta b$. The resulting closest consistent linear model $(A + \Delta A, b + \Delta b)$ is even more accurate than the LS one, since modifications of A are allowed.

Accuracy is the primary aim of LS and TLS, so it is not surprising that both solutions may exhibit very sensitive behavior to perturbations in the data matrices (A, b). Detailed sensitivity analyses for the LS and TLS problems may be found in [12, 18, 2, 44, 22, 14]. Many regularization methods have been proposed to decrease sensitivity and make LS and TLS applicable. Most regularization schemes for LS, including Tikhonov regularization [43], amount to solve a weighted LS problem for

* Received by the editors February 7, 1996; accepted for publication (in revised form) by S. Van Huffel November 4, 1996.
http://www.siam.org/journals/simax/18-4/29813.html
† Ecole Nationale Supérieure de Techniques Avancées, 32, Bd. Victor, 75739 Paris Cédex 15, France (elghaoui@ensta.fr, lebret@ensta.fr).

an augmented system. As pointed out in [18], the choice of weights (or regularization parameter) is usually not obvious and application dependent. Several criteria for optimizing the regularization parameter(s) have been proposed (see, e.g., [23, 11, 15]). These criteria are chosen according to some additional a priori information, of deterministic or stochastic nature. The extensive surveys [31, 8, 21] discuss these problems and some applications.

In contrast with the extensive work on sensitivity and regularization, relatively little has been done on the subject of *deterministic robustness* of LS problems in which the perturbations are deterministic and unknown but bounded (not necessarily small). Some work has been done on a qualitative analysis of the problem, where entries of (A, b) are unspecified except for their sign [26, 39]. In many papers mentioning least squares and robustness, the latter notion is understood in some stochastic sense; see, e.g., [20, 47, 37]. A notable exception concerns the field of identification, where the subject has been explored using a framework used in control system analysis [40, 9], or using regularization ideas combined with additional a priori information [34, 42].

In this paper, we assume that the data matrices are subject to (not necessarily small) deterministic perturbations. First, we assume that the given model is not a single pair (A, b) but a family of matrices $(A+\Delta A,\, b+\Delta b)$, where $\Delta = [\Delta A\ \Delta b]$ is an unknown-but-bounded matrix; precisely, $\|\Delta\| \leq \rho$, where $\rho \geq 0$ is given. For x fixed, we define the worst-case residual as

$$r(A, b, \rho, x) \triangleq \max_{\|\Delta A\ \Delta b\|_F \leq \rho} \|(A + \Delta A)x - (b + \Delta b)\|. \tag{1}$$

We say that x is a robust least squares (RLS) solution if x minimizes the worst-case residual $r(A, b, \rho, x)$. The RLS solution trades accuracy for robustness at the expense of introducing bias. In our paper, we assume that the perturbation bound ρ is known, but in section 3.5 we also show that TLS can be used as a preliminary step to obtain a value of ρ that is consistent with data matrices A, b.

In many applications, the perturbation matrices ΔA, Δb have a known structure. For instance, ΔA might have a Toeplitz structure inherited from A. In this case, the worst-case residual (1) might be a very conservative estimate. We are led to consider the following structured RLS (SRLS) problem. Given $A_0, \ldots, A_p \in \mathbf{R}^{n\times m}$, $b_0, \ldots, b_p \in \mathbf{R}^n$, we define, for every $\delta \in \mathbf{R}^p$,

$$\mathbf{A}(\delta) \triangleq A_0 + \sum_{i=1}^{p} \delta_i A_i, \quad \mathbf{b}(\delta) \triangleq b_0 + \sum_{i=1}^{p} \delta_i b_i. \tag{2}$$

For $\rho \geq 0$ and $x \in \mathbf{R}^m$, we define the structured worst-case residual as

$$r_S(\mathbf{A}, \mathbf{b}, \rho, x) \triangleq \max_{\|\delta\| \leq \rho} \|\mathbf{A}(\delta)x - \mathbf{b}(\delta)\|. \tag{3}$$

We say that x is an SRLS solution if x minimizes the worst-case residual $r_S(\mathbf{A}, \mathbf{b}, \rho, x)$.

Our main contribution is to show that we can compute the *exact value* of the optimal worst-case residuals using convex, second-order cone programming (SOCP) or semidefinite programming (SDP). The consequence is that the RLS and SRLS problems can be solved in polynomial time and with great practical efficiency using, e.g., recent interior-point methods [33, 46]. Our exact results are to be contrasted with those of Doyle et al. [9], who also use SDP to compute *upper bounds* on the worst-case residual for identification problems. In the preliminary draft [5] sent to us shortly

after submission of this paper, the authors provide a solution to an (unstructured) RLS problem, which is similar to that given in section 3.2.

Another contribution is to show that the RLS solution is continuous in the data matrices A, b. RLS can thus be interpreted as a (Tikhonov) regularization technique for ill-conditioned LS problems: the additional a priori information is ρ (the perturbation level), and the regularization parameter is optimal for robustness. Similar regularity results hold for the SRLS problem.

We also consider a generalization of the SRLS problem, referred to as the linear-fractional SRLS problem in what follows, in which the matrix functions $\mathbf{A}(\delta)$, $\mathbf{b}(\delta)$ in (2) depend rationally on the parameter vector δ. (We describe a robust interpolation problem that falls in this class in section 7.6.) Using the framework of [9], we show that the problem is NP-complete in this case, but we may compute and optimize upper bounds on the worst-case residual using SDP. In parallel with RLS, we interpret our solution as one of a weighted LS problem for an augmented system, the weights being computed via SDP.

The paper's outline is as follows. The next section is devoted to some technical lemmas. Section 3 is devoted to the RLS problem. In section 4, we consider the SRLS problem. Section 5 studies the linear-fractional SRLS problem. Regularity results are given in section 6. Section 7 shows numerical examples.

2. Preliminary results.

2.1. Semidefinite and second-order cone programs. We briefly recall some important results on semidefinite programs (SDPs) and second-order cone programs (SOCPs). These results can be found, e.g., in [4, 33, 46].

A linear matrix inequality is a constraint on a vector $x \in \mathbf{R}^m$ of the form

$$\mathcal{F}(x) = \mathcal{F}_0 + \sum_{i=1}^{m} x_i \mathcal{F}_i \geq 0, \tag{4}$$

where the symmetric matrices $\mathcal{F}_i = \mathcal{F}_i^T \in \mathbf{R}^{N\times N}$, $i = 0, \ldots, m$, are given. The minimization problem

$$\text{minimize } c^T x \text{ subject to } \mathcal{F}(x) \geq 0, \tag{5}$$

where $c \in \mathbf{R}^m$, is called an SDP. SDPs are convex optimization problems and can be solved in polynomial time with, e.g., primal-dual interior-point methods [33, 45].

The problem dual to problem (5) is

$$\begin{array}{ll} \text{maximize} & -\mathbf{Tr}\mathcal{F}_0\mathcal{Z} \\ \text{subject to} & \mathcal{Z} \geq 0, \ \ \mathbf{Tr}\mathcal{F}_i\mathcal{Z} = c_i, \ i = 1, \ldots, m, \end{array} \tag{6}$$

where $\mathcal{Z}$ is a symmetric $N \times N$ matrix and c_i is the ith coordinate of vector c. When both problems are strictly feasible (that is, when there exists $x, \mathcal{Z}$ which satisfy the constraints strictly), the existence of optimal points is guaranteed [33, Thm. 4.2.1], and both problems have equal optimal objectives. In this case, the optimal primal-dual pairs $(x, \mathcal{Z})$ are those pairs $(x, \mathcal{Z})$ such that x is feasible for the primal problem, $\mathcal{Z}$ is feasible for the dual one, and $\mathcal{F}(x)\mathcal{Z} = 0$.

An SOCP problem is one of the form

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \|C_i x + d_i\| \leq e_i^T x + f_i, \ \ i = 1, \ldots, L, \end{array} \tag{7}$$

where $C_i \in \mathbf{R}^{n_i \times m}$, $d_i \in \mathbf{R}^{n_i}$, $e_i \in \mathbf{R}^m$, $f_i \in \mathbf{R}$, $i = 1, \ldots, L$. The dual problem of problem (7) is

$$\begin{array}{ll} \text{maximize} & -\displaystyle\sum_{i=1}^{L} \left(d_i^T z_i + f_i s_i\right) \\ \text{subject to} & \displaystyle\sum_{i=1}^{L} \left(C_i^T z_i + e_i s_i\right) = c, \ \ \|z_i\| \le s_i, \ i = 1, \ldots, L, \end{array} \tag{8}$$

where $z_i \in \mathbf{R}^{n_i}$, $s_i \in \mathbf{R}$, $i = 1, \ldots, L$ are the dual variables. Optimality conditions similar to those for SDPs can be obtained for SOCPs. SOCPs can be expressed as SDPs; therefore, they can be solved in polynomial time using interior-point methods for SDPs. However, the SDP formulation is not the most efficient numerically, as special interior-point methods can be devised for SOCPs [33, 28, 1].

Precise complexity results on interior-point methods for SOCPs and SDPs are given by Nesterov and Nemirovsky [33, pp. 224, 236]. In practice, it is observed that the number of iterations is almost constant, independent of problem size [46]. For the SOCP, each iteration has complexity $O((n_1 + \cdots + n_L)m^2 + m^3)$; for the SDP, we refer the reader to [33].

2.2. $\mathcal{S}$-procedure. The following lemma can be found, e.g., in [4, p. 24]. It is widely used, e.g., in control theory and in connection with trust region methods in optimization [41].

LEMMA 2.1 ($\mathcal{S}$-procedure). *Let $F_0, \ldots, F_p$ be quadratic functions of the variable $\zeta \in \mathbf{R}^m$:*

$$F_i(\zeta) \stackrel{\Delta}{=} \zeta^T T_i \zeta + 2u_i^T \zeta + v_i, \ i = 0, \ldots, p,$$

where $T_i = T_i^T$. The following condition on $F_0, \ldots, F_p$:

$$F_0(\zeta) \ge 0 \text{ for all } \zeta \text{ such that } F_i(\zeta) \ge 0, \ i = 1, \ldots, p,$$

holds if

$$\text{there exist } \tau_1 \ge 0, \ldots, \tau_p \ge 0 \ \text{ such that } \begin{bmatrix} T_0 & u_0 \\ u_0^T & v_0 \end{bmatrix} - \sum_{i=1}^{p} \tau_i \begin{bmatrix} T_i & u_i \\ u_i^T & v_i \end{bmatrix} \ge 0.$$

When $p = 1$, the converse holds, provided that there is some ζ_0 such that $F_1(\zeta_0) > 0$.

The next lemma is a corollary of the above result in the case $p = 1$.

LEMMA 2.2. *Let $T_1 = T_1^T$, T_2, T_3, T_4 be real matrices of appropriate size. We have* $\det(I - T_4\Delta) \ne 0$ *and*

$$T(\Delta) = T_1 + T_2\Delta(I - T_4\Delta)^{-1}T_3 + T_3^T(I - T_4\Delta)^{-T}\Delta^T T_2^T \ge 0 \tag{9}$$

for every Δ, $\|\Delta\| \le 1$, if and only if $\|T_4\| < 1$ and there exists a scalar $\tau \ge 0$ such that

$$\begin{bmatrix} T_1 - \tau T_2 T_2^T & T_3^T - \tau T_2 T_4^T \\ T_3 - \tau T_4 T_2^T & \tau(I - T_4 T_4^T) \end{bmatrix} \ge 0. \tag{10}$$

Proof. If T_2 or T_3 equal zero, the result is obvious. Now assume $T_2, T_3 \neq 0$. Then, (10) implies $\tau > 0$, which in turn implies $\|T_4\| < 1$. Thus, for a given τ, (10) holds if and only if $\|T_4\| < 1$, and for every (u, p) we have

$$u^T(T_1 u + 2T_3^T p) - \tau(q^T q - p^T p) \geq 0,$$

where $q = T_2^T u + T_4^T p$. Since $T_2 \neq 0$, the constraint $q^T q \geq p^T p$ is qualified, that is, satisfied strictly for some (u_0, p_0) (choose $p_0 = 0$ and u_0 such that $T_2^T u_0 \neq 0$). Using the $\mathcal{S}$-procedure, we obtain that there exists $\tau \in \mathbf{R}$ such that (10) holds if and only if $\|T_4\| < 1$, and for every (u, p) such that $q^T q \geq p^T p$ we have $u^T(T_1 u + 2T_3^T p) \geq 0$. We end our proof by noting that for every pair (p, q), $p = \Delta^T q$ for some Δ, $\|\Delta\| \leq 1$ if and only if $p^T p \leq q^T q$. ∎

The following lemma is a "structured" version of the above, which can be traced back to [13].

LEMMA 2.3. *Let $T_1 = T_1^T$, T_2, T_3, T_4 be real matrices of appropriate size. Let $\mathcal{D}$ be a subspace of $\mathbf{R}^{N\times N}$ and denote by $\mathcal{S}$ (resp., $\mathcal{G}$) the set of symmetric (resp., skew-symmetric) matrices that commute with every element of $\mathcal{D}$. We have* $\det(I - T_4\Delta) \neq 0$ *and* (9) *for every $\Delta \in \mathcal{D}$, $\|\Delta\| \leq 1$, if there exist $S \in \mathcal{S}$, $G \in \mathcal{G}$ such that*

$$\begin{bmatrix} T_1 - T_2 S T_2^T & T_3^T - T_2 S T_4^T + T_2 G \\ T_3 - T_4 S T_2^T - G T_2^T & S - G T_4^T + T_4 G - T_4 S T_4^T \end{bmatrix} > 0, \quad S > 0.$$

If $\mathcal{D} = \mathbf{R}^{\mathcal{N}\times\mathcal{N}}$, the condition is necessary and sufficient.

Proof. The proof follows the scheme of that of Lemma 2.2, except that $p^T p \leq q^T q$ is replaced with $p^T S p \leq q^T S q$, $p^T G q = 0$, for given $S \in \mathcal{S}$, $S > 0$, $G \in \mathcal{G}$. Note that for $G = 0$, the above result is a simple application of Lemma 2.2 to the scaled matrices T_1, $T_2 S^{-1/2}$, $S^{1/2} T_3$, $S^{1/2} T_4 S^{-1/2}$. ∎

2.3. Elimination lemma. The last lemma is proven in [4, 24].

LEMMA 2.4 (elimination). *Given real matrices $W = W^T$, U, V of appropriate size, there exists a real matrix X such that*

$$W + UXV^T + VX^T U^T > 0 \tag{11}$$

if and only if

$$\tilde{U}^T W \tilde{U} > 0 \text{ and } \tilde{V}^T W \tilde{V} > 0, \tag{12}$$

where $\tilde{U}$, $\tilde{V}$ are orthogonal complements of U, V. If U, V are full column rank, and (12) *holds, a solution X to the inequality* (11) *is*

$$X = \sigma(U^T Q^{-1} U)^{-1} U^T Q^{-1} V, \tag{13}$$

where $Q \triangleq W + \sigma V V^T$, and σ is any scalar such that $Q > 0$ (the existence of which is guaranteed by (12)*).*

3. Unstructured RLS. In this section, we consider the RLS problem, which is to compute

$$\phi(A, b, \rho) \triangleq \min_x \max_{\|\Delta A \ \Delta b\|_F \leq \rho} \|(A + \Delta A)x - (b + \Delta b)\|. \tag{14}$$

For $\rho = 0$, we recover the standard LS problem. For every $\rho > 0$, $\phi(A, b, \rho) = \rho\phi(A/\rho, b/\rho, 1)$, so we take $\rho = 1$ in what follows, unless otherwise stated. In the remainder of this paper, $\phi(A, b)$ (resp., $r(A, b, x)$) denotes $\phi(A, b, 1)$ (resp., $r(A, b, 1, x)$).

In the preceding definition, the norm used for the perturbation bound is the Frobenius norm. As will be seen, the worst-case residual is the same when the norm used is the largest singular value norm.

3.1. Optimizing the worst-case residual. The following results yield a numerically efficient algorithm for solving the RLS problem in the unstructured case.

THEOREM 3.1. *When $\rho = 1$, the worst-case residual* (1) *is given by*

$$r(A, b, x) = \|Ax - b\| + \sqrt{\|x\|^2 + 1}.$$

The problem of minimizing $r(A, b, x)$ over $x \in \mathbf{R}^m$ has a unique solution x_{RLS}, referred to as the RLS solution. This problem can be formulated as the SOCP

$$\text{minimize } \lambda \text{ subject to } \|Ax - b\| \le \lambda - \tau, \quad \left\| \begin{bmatrix} x \\ 1 \end{bmatrix} \right\| \le \tau. \tag{15}$$

Proof. Fix $x \in \mathbf{R}^m$. Using the triangle inequality, we have

$$r(A, b, x) \le \|Ax - b\| + \sqrt{\|x\|^2 + 1}. \tag{16}$$

Now choose $\Delta = [\Delta A \ \Delta b]$ as

$$[\Delta A \ \Delta b] = \frac{u}{\sqrt{\|x\|^2 + 1}} \begin{bmatrix} x^T & 1 \end{bmatrix}, \text{ where } u = \begin{cases} \dfrac{Ax - b}{\|Ax - b\|} \text{ if } Ax \neq b, \\ \text{any unit-norm vector otherwise.} \end{cases}$$

Since Δ is rank one, we have $\|\Delta\|_F = \|\Delta\| = 1$. In addition, we have

$$\|(A + \Delta A)x - (b + \Delta b)\| = \|Ax - b\| + \sqrt{\|x\|^2 + 1},$$

which implies that Δ is a worst-case perturbation (for both the Frobenius and maximum singular value norms) and that equality always holds in (16). Finally, unicity of the minimizer x follows from the strict convexity of the worst-case residual. ∎

Using an interior-point primal-dual potential reduction method for solving the unstructured RLS problem (15), the number of iterations is almost constant [46]. Furthermore, each iteration takes $O((n + m)m^2)$ operations. A rough summary of this analysis is that the method has the same order of complexity as one singular value decomposition (SVD) of A.

3.2. Analysis of the optimal solution. Using duality results for SOCPs, we have the following theorem.

THEOREM 3.2. *When $\rho = 1$, the (unique) solution x_{RLS} to the RLS problem is given by*

$$x_{\text{RLS}} = \begin{cases} (\mu I + A^T A)^{-1} A^T b & \text{if } \mu \triangleq (\lambda - \tau)/\tau > 0, \\ A^\dagger b & \text{else,} \end{cases} \tag{17}$$

where (λ, τ) are the (unique) optimal points for problem (15).

Proof. Using the results of section 2.1, we obtain that the problem dual to (15) is

$$\text{maximize } b^T z - v \text{ subject to } A^T z + u = 0, \ \ \|z\| \le 1, \ \ \left\| \begin{bmatrix} u \\ v \end{bmatrix} \right\| \le 1.$$

Since both primal and dual problems are strictly feasible, there exist optimal points for both of them. If $\lambda = \tau$ at the optimum, then $Ax = b$, and

$$\lambda = \tau = \sqrt{\|x\|^2 + 1}.$$

In this case, the optimal x is the (unique) minimum-norm solution to $Ax = b$: $x = A^\dagger b$.

Now assume $\lambda > \tau$. Again, both primal and dual problems are strictly feasible; therefore, the primal- and dual-optimal objectives are equal:

$$\|Ax - b\| + \left\| \begin{bmatrix} x^T & 1 \end{bmatrix} \right\| = \lambda = b^T z - v = -(Ax - b)^T z - \begin{bmatrix} x^T & 1 \end{bmatrix} \begin{bmatrix} -A^T z \\ v \end{bmatrix}. \tag{18}$$

Using $\|z\| \le 1$, $\|[u^T \ v]^T\| \le 1$, $u = -A^T z$, we get

$$z = -\frac{Ax - b}{\|Ax - b\|} \text{ and } \begin{bmatrix} u^T & v \end{bmatrix} = -\frac{\begin{bmatrix} x^T & 1 \end{bmatrix}}{\sqrt{\|x\|^2 + 1}}.$$

Replace these values in $A^T z + u = 0$ to obtain the expression of the optimal x:

$$x = \left(A^T A + \mu I\right)^{-1} A^T b, \text{ with } \mu = \frac{\lambda - \tau}{\tau} = \frac{\|Ax - b\|}{\sqrt{\|x\|^2 + 1}}. \qquad \square$$

REMARK 3.1. *When $\lambda > \tau$, the RLS solution can be interpreted as the solution of a weighted LS problem for an augmented system:*

$$x_{\text{RLS}} = \arg\min \left\| \begin{bmatrix} A \\ I \\ 0 \end{bmatrix} x - \begin{bmatrix} b \\ 0 \\ 1 \end{bmatrix} \right\|_\Theta,$$

where $\Theta = \mathbf{diag}((\lambda - \tau)I, \tau I, \tau)$. The RLS method amounts to computing the weighting matrix Θ that is optimal for robustness via the SOCP (15). *We shall encounter a generalization of the above formula for the linear-fractional SRLS problem of section* 5.

REMARK 3.2. *It is possible to solve the problem when only A is perturbed ($\Delta b = 0$). In this case, the worst-case residual is $\|Ax - b\| + \|x\|$, and the optimal x is determined by* (17), *where $\mu\|x\| = \|Ax - b\|$. (See the example in section* 7.2.*)*

3.3. Reduction to a one-dimensional search. When the SVD of A is available, we can use it to reduce the problem to a one-dimensional convex differentiable problem. The following analysis will also be useful in section 6.

Introduce the SVD of A and a related decomposition for b:

$$A = U \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} V^T, \quad U^T b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

where $\Sigma = \mathbf{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbf{R}^{r \times r}$, $\Sigma > 0$, and $b_1 \in \mathbf{R}^r$, $r = \mathbf{Rank} A$.

Assume that $\lambda > \tau$ at the optimum of problem (15). From (18), we have

$$\begin{aligned} \lambda = b^T z - v &= \frac{b^T(b - Ax)}{\|Ax - b\|} + \frac{1}{\sqrt{\|x\|^2 + 1}} \\ &= \frac{1}{\tau} + \frac{b_2^T b_2}{\lambda - \tau} + b_1^T((\lambda - \tau)I + \tau\Sigma^2)^{-1} b_1. \end{aligned}$$

Since $\lambda = 0$ is never feasible, we may define $\theta = \tau/\lambda$. Multiplying by λ, we obtain that

$$\lambda^2 = \frac{1}{\theta} + \frac{b_2^T b_2}{1-\theta} + b_1^T((1-\theta)I + \theta\Sigma^2)^{-1} b_1.$$

From $\lambda \leq \|b\| + 1$ and $\tau \geq 1$, we deduce $\theta \geq \theta_{\min} \triangleq 1/(\|b\|+1)$. Thus, the optimal worst-case residual is

$$\phi(A,b)^2 = \inf_{\theta_{\min} \leq \theta < 1} f(\theta), \tag{19}$$

where f is the following function:

$$f(\theta) \triangleq \begin{cases} \dfrac{1}{\theta} + b^T((1-\theta)I + \theta AA^T)^{-1} b & \text{if } \theta_{\min} \leq \theta < 1, \\ \infty & \text{if } \theta = 1,\ b \notin \mathbf{Range}(A), \\ 1 + \|A^\dagger b\|^2 & \text{if } \theta = 1,\ b \in \mathbf{Range}(A). \end{cases} \tag{20}$$

The function f is convex and twice differentiable on $[\theta_{\min}\ 1[$. If $b \notin \mathbf{Range}(A)$, f is infinite at $\theta = 1$; otherwise, f is twice differentiable on the closed interval $[\theta_{\min}\ 1]$. Therefore, the minimization of f can be done using standard Newton methods for differentiable optimization.

THEOREM 3.3. *When $\rho = 1$, the solution of the unstructured RLS can be computed by solving the one-dimensional convex differentiable problem* (19) *or by computing the unique real root inside $[\theta_{\min}\ 1]$ (if any) of the equation*

$$\frac{1}{\theta^2} = \frac{\|b_2\|^2}{(1-\theta)^2} + \sum_{i=1}^{r} \frac{b_{1i}^2(1-\sigma_i^2)}{(1+\theta(\sigma_i^2-1))^2}.$$

The above theorem yields an alternative method for computing the RLS solution. This method is similar to the one given in [5]. A related approach was used for quadratically constrained LS problems in [19].

The above solution, which requires one SVD of A, has cost $O(nm^2 + m^3)$. The SOCP method is only a few times more costly (see the end of section 3.1), with the advantage that we can include all kinds of additional constraints on x (nonnegativity and/or quadratic constraints, etc.) in the SOCP (15), with low additional cost. Also, the SVD solution does not extend to the structured case considered in section 4.

3.4. Robustness of LS solution. It is instructive to know when the RLS and LS solutions coincide, in which case we can say the LS solution is robust. This happens if and only if the optimal θ in problem (19) is equal to 1. The latter implies $b_2 = 0$ (that is, $b \in \mathbf{Range}(A)$). In this case, f is differentiable at $\theta = 1$, and its minimum over $[\theta_{\min}\ 1]$ is at $\theta = 1$ if and only if

$$\frac{df}{d\theta}(1) = b_1^T \Sigma^{-4} b_1 - (1 + b_1^T \Sigma^{-2} b_1) \leq 0.$$

We obtain a necessary and sufficient condition for the optimal θ to be equal to 1. This condition is

$$b \in \mathbf{Range}(A),\ \ b^T (AA^T)^{2\dagger} b \leq 1 + b^T (AA^T)^\dagger b. \tag{21}$$

If (21) holds, then the RLS and LS solutions coincide. Otherwise, the optimal $\theta < 1$, and x is given by (17). We may write the latter condition in the case when the norm bound of the perturbation ρ is different from 1 as the following: $\rho > \rho_{\min}$, where

$$\rho_{\min}(A,b) \triangleq \begin{cases} \dfrac{\sqrt{1+\|A^\dagger b\|^2}}{\|(AA^T)^\dagger b\|} & \text{if } b \in \mathbf{Range}(A),\ A \neq 0,\ b \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

Thus, $\rho_{\min}$ can be interpreted as the perturbation level that the LS solution allows. We note that when $b \in \mathbf{Range}(A)$, the LS and TLS solutions also coincide.

COROLLARY 3.4. *The LS, TLS, and RLS solutions coincide whenever the norm bound on the perturbation matrix ρ satisfies $\rho \leq \rho_{\min}(A,b)$, where $\rho_{\min}(A,b)$ is defined in* (22). *Thus, $\rho_{\min}(A,b)$ can be seen as a robustness measure of the LS (or TLS) solution.*

When A is full rank, the robustness measure $\rho_{\min}$ is nonzero and decreases as the condition number of A increases.

REMARK 3.3. *We note that the TLS solution x_{TLS} is the most accurate, in the sense that it minimizes the distance function (see* [18]*)*

$$a(A,b,x) = \frac{\|Ax-b\|}{\sqrt{\|x\|^2+1}},$$

and is the least robust, in the sense of the worst-case residual. The LS solution, $x_{\text{LS}} = A^\dagger b$, is intermediate (in the sense of accuracy and robustness). In fact, it can be shown that

$$\begin{gathered} r(A,b,x_{\text{RLS}},\rho) \leq r(A,b,x_{\text{LS}},\rho) \leq r(A,b,x_{\text{TLS}},\rho), \\ a(A,b,x_{\text{TLS}}) \leq a(A,b,x_{\text{LS}}) \leq a(A,b,x_{\text{RLS}}), \\ \|x_{\text{RLS}}\| \leq \|x_{\text{LS}}\| \leq \|x_{\text{TLS}}\|. \end{gathered}$$

3.5. RLS and TLS. The RLS framework assumes that the data matrices (A,b) are the "nominal" values of the model, which are subject to unstructured perturbation, bounded in norm by ρ. Now, if we think of (A,b) as "measured" data, the assumption that (A,b) correspond to a nominal model may not be judicious. Also, in some applications, the norm bound ρ on the perturbation may be hard to estimate. The TLS solution, when it exists, can be used in conjunction with RLS to address this issue.

Assume that the TLS problem has a solution. Let ΔA_{TLS}, Δb_{TLS}, x_{TLS} be minimizers of the TLS problem

$$\text{minimize } \|\Delta A\ \Delta b\|_F \text{ subject to } (A+\Delta A)x = b + \Delta b,$$

and let

$$\rho_{\text{TLS}} = \|\Delta A_{\text{TLS}}\ \Delta b_{\text{TLS}}\|_F, \quad A_{\text{TLS}} = A + \Delta A_{\text{TLS}}, \quad b_{\text{TLS}} = A + \Delta b_{\text{TLS}}.$$

TLS finds a consistent, linear system that is closest (in Frobenius norm sense) to the observed data (A,b). The underlying assumption is that the observed data (A,b) is the result of a consistent, linear system which, under the measurement process, has been subjected to unstructured perturbations, unknown but bounded in norm by ρ_{TLS}. With this assumption, any point of the ball

$$\{(A',b') \mid \|A' - A_{\text{TLS}}\ b' - b_{\text{TLS}}\|_F \leq \rho_{\text{TLS}}\}$$

can be observed, just as well as (A, b). Thus, TLS computes an "uncertain linear system" representation of the observed phenomenon: $(A_{\rm TLS}, b_{\rm TLS})$ is the nominal model, and $\rho_{\rm TLS}$ is the perturbation level.

Once this uncertain system representation ($A_{\rm TLS}$, $b_{\rm TLS}$, $\rho_{\rm TLS}$) is computed, choosing $x_{\rm TLS}$ as a "solution" to $Ax \simeq b$ amounts to finding the exact solution to the nominal system. Doing so, we compute a very accurate solution (with zero residual), which does not take into account the perturbation level $\rho_{\rm TLS}$. A more robust solution is given by the solution to the following RLS problem:

$$\min_x \max_{\|\Delta A\ \Delta b\|_F \le \rho_{\rm TLS}} \|(A_{\rm TLS} + \Delta A)x - (b_{\rm TLS} + \Delta b)\|. \tag{23}$$

The solution to the above problem coincides with the TLS one (that is, in our case, with $x_{\rm TLS}$) when $\rho_{\rm TLS} \le \rho_{\min}(A_{\rm TLS}, b_{\rm TLS})$. (Since $b_{\rm TLS} \in \mathbf{Range}(A_{\rm TLS})$, the latter quantity is strictly positive, except when $A_{\rm TLS} = 0$, $b_{\rm TLS} = 0$.)

With standard LS, the perturbations that account for measurement errors are structured (with $\Delta A = 0$). To be consistent with LS, one should consider the following RLS problem instead of (23):

$$\min_x \max_{\|\Delta b\| \le \rho_{\rm LS}} \|A_{\rm LS}x - (b_{\rm LS} + \Delta b)\|. \tag{24}$$

It turns out that the above problem yields the same solution as LS itself.

To summarize, RLS can be used in conjunction with TLS for "solving" a linear system $Ax \simeq b$. Solve the TLS problem to build an "uncertain linear system" representation ($A_{\rm TLS}$, $b_{\rm TLS}$, $\rho_{\rm TLS}$) of the observed data. Then, take the solution $x_{\rm RLS}$ to the RLS problem with the nominal matrices $(A_{\rm TLS}, b_{\rm TLS})$, and uncertainty size $\rho_{\rm TLS}$. Note that computing the TLS solution (precisely, $A_{\rm TLS}$, $b_{\rm TLS}$, and $\rho_{\rm TLS}$) only requires the computation of the smallest singular value and associated singular subspace [17].

4. Structured Robust Least Squares (SRLS). In this section, we consider the SRLS problem, which is to compute

$$\phi_S(\mathbf{A}, \mathbf{b}, \rho) \triangleq \min_x \max_{\|\delta\| \le \rho} \|\mathbf{A}(\delta)x - \mathbf{b}(\delta)\|, \tag{25}$$

where $\mathbf{A}, \mathbf{b}$ are defined in (2). As before, we assume with no loss of generality that $\rho = 1$ and denote $r_S(\mathbf{A}, \mathbf{b}, 1, x)$ by $r_S(\mathbf{A}, \mathbf{b}, x)$. Throughout the section, we use the following notation:

$$M(x) \triangleq \left[\begin{array}{ccc} A_1x - b_1 & \dots & A_px - b_p \end{array}\right]. \tag{26}$$

4.1. Computing the worst-case residual. We first examine the problem of computing the worst-case residual $r_S(\mathbf{A}, \mathbf{b}, x)$ for a given $x \in \mathbf{R}^m$. Define

$$F \triangleq M(x)^T M(x), \quad g \triangleq M(x)^T(A_0x - b_0), \quad h \triangleq \|A_0x - b_0\|^2. \tag{27}$$

With the above notation, we have

$$r_S(\mathbf{A}, \mathbf{b}, x)^2 = \max_{\delta^T\delta \le 1} \left[\begin{array}{c} 1 \\ \delta \end{array}\right] \left[\begin{array}{cc} h & g^T \\ g & F \end{array}\right] \left[\begin{array}{c} 1 \\ \delta \end{array}\right]. \tag{28}$$

Now let $\lambda \ge 0$. Using the $\mathcal{S}$-procedure (Lemma 2.1), we have

$$\left[\begin{array}{c} 1 \\ \delta \end{array}\right] \left[\begin{array}{cc} h & g^T \\ g & F \end{array}\right] \left[\begin{array}{c} 1 \\ \delta \end{array}\right] \le \lambda$$

for every δ, $\delta^T\delta \le 1$ if and only if there exists a scalar $\tau \ge 0$ such that

$$\begin{bmatrix} 1 \\ \delta \end{bmatrix} \begin{bmatrix} \lambda - \tau - h & -g^T \\ -g & \tau I - F \end{bmatrix} \begin{bmatrix} 1 \\ \delta \end{bmatrix} \ge 0 \text{ for every } \delta \in \mathbf{R}^p.$$

Using the fact that $\tau \ge 0$ is implied by $\tau I \ge F$, we may rewrite the above condition as

$$\mathcal{F}(\lambda, \tau) \triangleq \begin{bmatrix} \lambda - \tau - h & -g^T \\ -g & \tau I - F \end{bmatrix} \ge 0. \tag{29}$$

The consequence is that the worst-case residual is computed by solving an SDP with two scalar variables. A bit more analysis shows how to reduce the problem to a one-dimensional, convex differentiable problem and how to obtain the corresponding worst-case perturbation.

THEOREM 4.1. *For every x fixed, the squared worst-case residual (for $\rho = 1$) $r_S(\mathbf{A}, \mathbf{b}, x)^2$ can be computed by solving the SDP in two variables*

$$\textit{minimize } \lambda \textit{ subject to } (29),$$

or, alternatively, by minimizing a one-dimensional convex differentiable function

$$r_S(\mathbf{A}, \mathbf{b}, x)^2 = h + \inf_{\tau \ge \lambda_{\max}(F)} f(\tau), \tag{30}$$

where

$$f(\tau) \triangleq \begin{cases} \tau + g^T(\tau I - F)^{-1} g & \textit{if } \tau > \lambda_{\max}(F), \\ \infty & \textit{if } \tau = \lambda_{\max}(F) \textit{ is } (F,g)\textit{-controllable}, \\ \lambda_{\max}(F) + g^T(\tau I - F)^{\dagger} g & \textit{if } \tau = \lambda_{\max}(F) \textit{ is not } (F,g)\textit{-controllable}. \end{cases} \tag{31}$$

If τ is optimal for problem (30)*, the equations in δ*

$$(\tau I - F)\delta = g, \quad \|\delta\| = 1$$

have a solution, any of which is a worst-case perturbation.

Proof. See Appendix A, where we also show how to compute a worst-case perturbation. □

4.2. Optimizing the worst-case residual. Using Theorem 4.1, the expression of F, g, h given in (27), and Schur complements, we obtain the following result.

THEOREM 4.2. *When $\rho = 1$, the Euclidean-norm SRLS can be solved by computing an optimal solution (λ, τ, x) of the SDP*

$$\textit{minimize } \lambda \textit{ subject to } \begin{bmatrix} \lambda - \tau & 0 & (A_0 x - b_0)^T \\ 0 & \tau I & M(x)^T \\ A_0 x - b_0 & M(x) & I \end{bmatrix} \ge 0, \tag{32}$$

where $M(x)$ is defined in (26)*.*

REMARK 4.1. *Straightforward manipulations show that the results are coherent with the unstructured case.*

Although the above SDP is not directly amenable to the more efficient SOCP formulation, we may devise special interior-point methods for solving the problem. These special-purpose methods will probably have much greater efficiency than general-purpose SDP solvers. This study is left for the future.

REMARK 4.2. *The discussion of section* 3.5 *extends to the case when the perturbations are structured. TLS problems with (affine) structure constraints on perturbation matrices are discussed in* [7]*. While the structured version of the TLS problem becomes very hard to solve, the SRLS problem retains polynomial-time complexity.*

5. Linear-fractional SRLS. In this section, we examine a generalization of the SRLS problem. Our framework encompasses the case when the functions $\mathbf{A}(\delta)$, $\mathbf{b}(\delta)$ are rational. We show that the computation of the worst-case residual is NP-complete but that upper bounds can be computed (and optimized) using SDP. First, we need to motivate the problem and develop a formalism for posing it. This formalism was introduced by Doyle et al. [9] in the context of robust identification.

5.1. Motivations. In some structured robust least-squares problems such as (3), it may not be convenient to measure the perturbation size with Euclidean norm. Indeed, the latter implies a correlated bound on the perturbation. One may instead consider an SRLS problem in which the bounds are not correlated; that is, the perturbation size in (3) is measured by the maximum norm

$$\min_x \max_{\|\delta\|_\infty \le 1} \|\mathbf{A}(\delta)x - \mathbf{b}(\delta)\|. \tag{33}$$

Also, in some RLS problems, we may assume that some columns of $[A\ b]$ are perfectly known. For instance, the error $[\Delta A\ \Delta b]$ has the form $[\Delta A\ 0]$, where ΔA is bounded and otherwise unknown. More generally, we may be interested in SRLS problems, where the perturbed data matrices write

$$\left[\begin{array}{cc} \mathbf{A}(\Delta) & \mathbf{b}(\Delta) \end{array}\right] = \left[\begin{array}{cc} A & b \end{array}\right] + L\Delta \left[\begin{array}{cc} R_A & R_b \end{array}\right], \tag{34}$$

where A, b, L, R_A, R_b are given matrices, and Δ is a (full) norm-bounded matrix. In such a problem, the perturbation is not structured, except via the matrices L, R_A, R_b. (Note that a special case of this problem is solved in [5].)

Finally, we may be interested in SRLS problems in which the matrix functions $\mathbf{A}(\delta)$, $\mathbf{b}(\delta)$ in (3) are rational functions of the parameter vector δ. One example is given in section 7.6.

It turns out that the extensions described in the three preceding paragraphs can be addressed using the same formalism, which we now detail.

5.2. Problem definition. Let $\mathcal{D}$ be a subspace of $\mathbf{R}^{N\times N}$, $A \in \mathbf{R}^{n\times m}$, $b \in \mathbf{R}^n$, $L \in \mathbf{R}^{n\times N}$, $R_A \in \mathbf{R}^{N\times m}$, $R_b \in \mathbf{R}^N$, $D \in \mathbf{R}^{N\times N}$. For every $\Delta \in \mathcal{D}$ such that $\det(I - D\Delta) \neq 0$, we define the matrix functions

$$\left[\begin{array}{cc} \mathbf{A}(\Delta) & \mathbf{b}(\Delta) \end{array}\right] = \left[\begin{array}{cc} A & b \end{array}\right] + L\Delta(I - D\Delta)^{-1} \left[\begin{array}{cc} R_A & R_b \end{array}\right].$$

For a given $x \in \mathbf{R}^m$, we define the worst-case residual by

$$r_{\mathcal{D}}(\mathbf{A}, \mathbf{b}, \rho, x) \triangleq \left\{ \begin{array}{ll} \max\limits_{\Delta\in\mathcal{D},\ \|\Delta\|\le\rho} \|\mathbf{A}(\Delta)x - \mathbf{b}(\Delta)\| & \text{if } \det(I - D\Delta) \neq 0, \\ \infty & \text{else.} \end{array} \right. \tag{35}$$

We say that x is an SRLS solution if x minimizes the worst-case residual above. As before, we assume $\rho = 1$ with no loss of generality and denote $r_{\mathcal{D}}(\mathbf{A}, \mathbf{b}, 1, x)$ by $r_{\mathcal{D}}(\mathbf{A}, \mathbf{b}, x)$.

The above formulation encompasses the three situations referred to in section 5.1. First, the maximum-norm SRLS problem (33) is readily transformed into problem (35) as follows. Let $L_i \in \mathbf{R}^{n\times N}$, $R_i \in \mathbf{R}^{N\times(m+1)}$ be such that $[A_i\ b_i] = L_i R_i$, $\mathbf{Rank} L_i = \mathbf{Rank} R_i = r_i$, where $r_i = \mathbf{Rank}[A_i\ b_i]$. Set $D = 0$, and let

$$\begin{array}{c} L = \left[\begin{array}{ccc} L_1 & \dots & L_p \end{array}\right], \quad R^T = \left[\begin{array}{ccc} R_1^T & \dots & R_p^T \end{array}\right], \\ \mathcal{D} = \left\{ \bigoplus_{i=1}^p \delta_i I_{s_i} \ \mid \delta_i \in \mathbf{R},\ 1 \le i \le p \right\}. \end{array} \tag{36}$$

Problem (33) can be formulated as the minimization of (35), with $\mathcal{D}$ defined as above.

Also, we recover the case when the perturbed matrices write as in (34) when we allow Δ to be any full matrix (that is, $\mathcal{D} = \mathbf{R}^{N\times N}$). In particular, we recover the unstructured RLS problem of section 3 as follows. Assume $n > m$. We have

$$\left[\begin{array}{cc} \Delta A & \Delta b \end{array}\right] = L \left[\begin{array}{ccc} \Delta A & \Delta b & \times \end{array}\right] R,$$

where $L = I$, $R^T = [I\ 0]$. (The symbol $\times$ refers to dummy elements that are added to the perturbation matrix in order to make it a square, $n \times n$ matrix.) In this case, the perturbation set $\mathcal{D}$ is $\mathbf{R}^{n\times n}$.

Finally, the case when $\mathbf{A}(\delta)$ and $\mathbf{b}(\delta)$ are rational functions of a vector δ (well defined over the unit ball $\{\delta \mid \|\delta\|_\infty \leq 1\}$) can be converted (in polynomial time) into the above framework (see, e.g., [48] for a conversion procedure). We give an example of such a conversion in section 7.6.

5.3. Complexity analysis. In comparison with the SRLS problem of section 4, the linear-fractional SRLS problem offers two levels of increased complexity.

First, checking whether the worst-case residual is finite is NP-complete [6]. The linear-fractional dependence (that is, $D \neq 0$) is a first cause of increased complexity.

The SRLS problem above remains hard even when matrices $\mathbf{A}(\delta)$, $\mathbf{b}(\delta)$ depend affinely on the perturbation elements ($D = 0$). Consider, for instance, the SRLS problem with $D = 0$ and in which $\mathcal{D}$ is defined as in (36). In this case, the problem of computing the worst-case residual can be formulated as

$$\max_{\|\delta\|_\infty \leq 1} \left[\begin{array}{c} 1 \\ \delta \end{array}\right] \left[\begin{array}{cc} h & g^T \\ g & F \end{array}\right] \left[\begin{array}{c} 1 \\ \delta \end{array}\right]$$

for appropriate F, g, h. The only difference with the worst-case residual defined in (28) is the norm used to measure perturbation. Computing the above quantity is NP-complete (it is equivalent to a MAX CUT problem [36, 38]). The following lemma, which we provide for the sake of completeness, is a simple corollary of a result by Nemirovsky [32].

LEMMA 5.1. *Consider the problem $\mathcal{P}(\mathbf{A}, \mathbf{b}, \mathcal{D}, x)$ defined as follows: given a positive rational number λ, matrices A, b, L, R_A, R_b, D of appropriate size, and an m-vector x, all with rational entries, and a linear subset $\mathcal{D}$, determine whether $r_{\mathcal{D}}(\mathbf{A}, \mathbf{b}, x) \leq \lambda$. Problem $\mathcal{P}(\mathbf{A}, \mathbf{b}, \mathcal{D}, x)$ is NP-complete.*

Proof. See Appendix B. □

5.4. An upper bound on the worst-case residual. Although our problem is NP-complete, we can minimize upper bounds in polynomial time using SDP. Introduce the following linear subspaces:

$$\begin{aligned} &\mathcal{B} \triangleq \{B \in \mathbf{R}^{N\times N} \mid B\Delta = \Delta B \text{ for every } \Delta \in \mathcal{D}\}, \\ &\mathcal{S} \triangleq \left\{S \in \mathcal{B} \;\middle|\; S = S^T\right\},\ \mathcal{G} \triangleq \left\{G \in \mathcal{B} \;\middle|\; G = -G^T\right\}. \end{aligned} \tag{37}$$

Let $\lambda \in \mathbf{R}$. The inequality $\lambda > r_{\mathcal{D}}(\mathbf{A}, \mathbf{b}, x)$ holds if and only if, for every $\Delta \in \mathcal{D}$, $\|\Delta\| \leq 1$, we have $\det(I - D\Delta) \neq 0$ and

$$\left[\begin{array}{cc} \lambda I & Ax - b \\ (Ax-b)^T & \lambda \end{array}\right] + \left[\begin{array}{c} L \\ 0 \end{array}\right] \Delta (I - D\Delta)^{-1} \left[\begin{array}{cc} 0 & R_A x - R_b \end{array}\right]$$
$$+ \left[\begin{array}{c} 0 \\ (R_A x - R_b)^T \end{array}\right] (I - D\Delta)^{-T} \Delta^T \left[\begin{array}{cc} L^T & 0 \end{array}\right] > 0.$$

Using Lemma 2.3, we obtain that $\lambda > r_{\mathcal{D}}(\mathbf{A}, \mathbf{b}, x)$ holds if there exist $S \in \mathcal{S}$, $G \in \mathcal{G}$, such that

$$\mathcal{F}(\lambda, S, G, x) = \left[\begin{array}{cc|c} \multicolumn{2}{c|}{\Theta} & \begin{array}{c} Ax - b \\ R_A x - R_b \end{array} \\ \hline (Ax-b)^T & (R_A x - R_b)^T & \lambda \end{array}\right] > 0, \tag{38}$$

where

$$\Theta \triangleq \left[\begin{array}{cc} \lambda I - LSL^T & -LSD^T + LG \\ -DSL^T + G^T L^T & S + DG - GD^T - DSD^T \end{array}\right]. \tag{39}$$

Minimizing λ subject to the above semidefinite constraint yields an upper bound for $r_{\mathcal{D}}(\mathbf{A}, \mathbf{b}, x)$. It turns out that the above estimate of the worst-case residual is actually *exact* in some "generic" sense.

THEOREM 5.2. *When $\rho = 1$, an upper bound on the worst-case residual $r_{\mathcal{D}}(\mathbf{A}, \mathbf{b}, x)$ can be obtained by solving the SDP*

$$\inf_{S,G,\lambda} \lambda \text{ subject to } S \in \mathcal{S},\ G \in \mathcal{G},\ (38). \tag{40}$$

The upper bound is exact when $\mathcal{D} = \mathbf{R}^{N \times N}$. If $\Theta > 0$ at the optimum, the upper bound is also exact.

Proof. See Appendix C. □

5.5. Optimizing the worst-case residual. Since x appears linearly in the constraint (38), we may optimize the worst-case residual's upper bound using SDP. We may reduce the number of variables appearing in the previous problem, using the elimination Lemma 2.4. Inequality in (38) can be written as in (11) with

$$W = \left[\begin{array}{cc|c} \multicolumn{2}{c|}{\Theta} & \begin{array}{c} -b \\ -R_b \end{array} \\ \hline -b & -R_b & \lambda \end{array}\right], \quad U = \left[\begin{array}{c} A \\ R_A \\ 0 \end{array}\right], \quad V = \left[\begin{array}{c} 0 \\ 0 \\ 1 \end{array}\right],$$

where Θ is defined in (39).

Denote by $\mathcal{N}$ the orthogonal complement of $[A^T \ R_A^T]^T$. Using the elimination Lemma 2.4, we obtain an equivalent condition for (38) to hold for some $x \in \mathbf{R}^m$; namely,

$$S \in \mathcal{S},\ G \in \mathcal{G},\ \Theta > 0,\ (\mathcal{N} \bigoplus 1)^T \left[\begin{array}{cc|c} \multicolumn{2}{c|}{\Theta} & \begin{array}{c} -b \\ -R_b \end{array} \\ \hline -b & -R_b & \lambda \end{array}\right] (\mathcal{N} \bigoplus 1) > 0. \tag{41}$$

For every λ, S, G that are strictly feasible for the above constraints, an x that satisfies (38) is given, when R_A is full rank, by

$$x = \left(\left[\begin{array}{cc} A^T & R_A^T \end{array}\right] \Theta^{-1} \left[\begin{array}{c} A \\ R_A \end{array}\right] \right)^{-1} \left[\begin{array}{cc} A^T & R_A^T \end{array}\right] \Theta^{-1} \left[\begin{array}{c} b \\ R_b \end{array}\right]. \tag{42}$$

(To prove this, we applied formula (13) and took $\sigma \to \infty$.)

THEOREM 5.3. *When $\rho = 1$, an upper bound on the optimal worst-case residual can be obtained by solving the SDP*

$$\inf_{S,G,\lambda,x} \lambda \text{ subject to } S \in \mathcal{S},\ G \in \mathcal{G},\ (38), \tag{43}$$

or, alternatively, the SDP

$$\inf_{S,G,\lambda} \lambda \text{ subject to } (41). \tag{44}$$

The upper bound is always exact when $\mathcal{D} = \mathbf{R}^{N\times N}$. *If* $\Theta > 0$ *at the optimum, the upper bound is also exact. The optimal* x *is then unique and given by* (42) *when* R_A *is full rank.*

Proof. See Appendix C. □

REMARK 5.1. *In parallel to the unstructured case (see Remark* 3.1*), the linear-fractional SRLS can be interpreted as a weighted LS for an augmented system. Precisely, when* $\Theta > 0$, *the linear-fractional SRLS solution can be interpreted as the solution of a weighted LS problem*

$$x_{\text{SRLS}} \in \arg\min \left\| \begin{bmatrix} A \\ R_A \end{bmatrix} x - \begin{bmatrix} b \\ R_b \end{bmatrix} \right\|_\Theta .$$

The SRLS method amounts to computing the weighting matrix Θ *that is optimal for robustness.*

REMARK 5.2. *Our results are coherent with the unstructured case: replace* L *by* I, R *by* $[I\ 0]^T$, *variable* S *by* τI, *and set* $G = 0$. *The parameter* μ *of Theorem* 3.2 *can be interpreted as the Schur complement of* $\lambda I - LSL^T$ *in the matrix* Θ.

REMARK 5.3. *We emphasize that the above results are exact (nonconservative) when the perturbation structure is full. In particular, we recover (and generalize) the results of* [5] *in the case when only some columns of* A *are affected by otherwise unstructured perturbations.*

REMARK 5.4. *When* $D = 0$, *it is possible to use the approximation method of* [16] *to obtain solutions (based on the SDP relaxations given in Theorem* 5.3*) that have expected value within* 14% *of the true value.*

6. Link with regularization. The standard LS solution x_{LS} is very sensitive to errors in A, b when A is ill conditioned. In fact, the LS solution might not be a continuous function of A, b when A is near deficient. This has motivated many researchers to look for ways to regularize the LS problem, which is to make the solution x unique and continuous in the data matrices (A, b). In this section, we briefly examine the links of our RLS and SRLS solution with regularization methods for standard LS.

Beforehand, we note that since all our problems are formulated as SDPs, we could invoke the quite complete sensitivity analysis results obtained by Bonnans, Cominetti, and Shapiro [3]. The application of these general results to our SDPs is considered in [35].

6.1. Regularization methods for LS. Most regularization methods for LS require imposing an additional bound on the solution vector x. One way is to minimize $\|Ax - b\|^2 + \Omega(x)$, where Ω is some squared norm (see [23, 43, 8]). Another way is to use constrained least squares (see [18, pp. 561–571]).

In a classical Tikhonov regularization method, $\Omega(x) = \mu\|x\|^2$, where $\mu > 0$ is some "regularization" parameter. The modified value of x is obtained by solving an augmented LS problem

$$\text{minimize } \|Ax - b\|^2 + \mu\|x\|^2 \tag{45}$$

and is given by

$$x(\mu) = (\mu I + A^T A)^{-1} A^T b. \tag{46}$$

(Note that for every $\mu > 0$, the above x is continuous in (A, b).)

The above expression also arises in the Levenberg–Marquardt method for optimization or in the Ridge regression problem [17]. As mentioned in [18], the choice of an appropriate μ is problem dependent and in many cases not obvious.

In more elaborate regularization schemes of the Tikhonov type, the identity matrix in (46) is replaced with a positive semidefinite weighting matrix (see for instance [31, 8]). Again, this can be interpreted as a (weighted) least-squares method for an augmented system.

6.2. RLS and regularization. Noting the similarity between (17) and (46), we can interpret the (unstructured) RLS method as that of Tikhonov regularization. The following theorem yields an estimate of the "smoothing effect" of the RLS method. Note that improved regularity results are given in [35].

THEOREM 6.1. *The (unique) RLS solution x_{RLS} and the optimal worst-case residual are continuous functions of the data matrices A, b. Furthermore, if $\mathcal{K}$ is a compact set of $\mathbf{R}^n$, and $d_{\mathcal{K}} = \max\{\|b\| \mid b \in \mathcal{K}\}$, then for every uncertainty size $\rho > 0$, the function*

$$\begin{array}{ccc} \mathbf{R}^{n\times m} \times \mathcal{K} & \longrightarrow & [1 \;\; d_{\mathcal{K}} + 1], \\ (A, b) & \longmapsto & \phi(A, b, \rho) \end{array}$$

is Lipschitzian, with Lipschitz constant $1 + d_{\mathcal{K}}/\rho$.

Theorem 6.1 shows that any level of robustness (that is, any norm bound on perturbations $\rho > 0$) guarantees regularization. We describe in section 7 some numerical examples that illustrate our results.

REMARK 6.1. *In the RLS method, the Tikhonov regularization parameter μ is chosen by solving a second-order cone problem in such a way that μ is optimal for* robustness. *The cost of the RLS solution is equal to the cost of solving a small number of least-squares problems of the same size as the classical Tikhonov regularization problem* (45).

REMARK 6.2. *The equation that determines μ in the RLS method is*

$$\mu = \frac{\|Ax(\mu) - b\|}{\rho\sqrt{\|x(\mu)\|^2 + 1}}.$$

This choice resembles Miller's choice [30], *where μ is determined recursively by the equations*

$$\mu = \frac{\|Ax(\mu) - b\|}{\rho\|x(\mu)\|}.$$

This formula arises in RLS when there is no perturbation in b (see Remark 3.2*). Thus, Miller's solution corresponds to an RLS problem in which the perturbation affects only the columns of A. We note that this solution is not necessarily regular (continuous).*

TLS deserves a special mention here. When the TLS problem has a solution, it is given by $x_{\text{TLS}} = (A^T A - \sigma^2 I)^{-1} A^T b$, where σ is the smallest singular value of $[A \; b]$. This corresponds to $\mu = -\sigma^2$ in (46). The negative value of μ implies that the TLS is a "deregularized" LS, a fact noted in [17]. In view of our link between regularization and robustness, the above is consistent with the fact that RLS trades off the accuracy of TLS with robustness and regularity, at the expense of introducing bias in the solution. See also Remark 3.3.

6.3. SRLS and regularization. Similarly, we may ask whether the solution to the SRLS problem of section 4 is continuous in the data matrices A_i, b_i, as was the case for unstructured RLS problems. We only discuss continuity of the optimal worst-case residual with respect to (A_0, b_0) (in many problems, the coefficient matrices A_i, b_i for $i = 1, \ldots, p$ are fixed).

In view of Theorem 4.2, continuity holds if the feasible set of the SDP (32) is bounded. Obviously, the objective λ is bounded above by

$$\max_{\delta^T\delta \le 1} \left\| b_0 + \sum_{i=1}^{p} \delta_i b_i \right\| \le \|b_0\| + \sum_{i=1}^{p} \|b_i\|.$$

Thus the variable τ is also bounded, as (32) implies $0 \le \tau \le \lambda$. With λ, τ bounded above, we see that (32) implies that x is bounded if

$$\left\| \begin{array}{cccc} A_0x - b_0 & A_1x - b_1 & \ldots & A_px - b_p \end{array} \right\| \text{ bounded implies } x \text{ bounded.}$$

The above property holds if and only if $[A_0^T\ A_1^T \ldots A_p^T]^T$ is full rank.

THEOREM 6.2. *A sufficient condition for continuity of the optimal worst-case residual (as a function of (A_0, b_0)) is that $[A_1^T \ldots A_p^T]^T$ is full rank.*

6.4. Linear-fractional SRLS and regularization. Precise conditions for continuity of the optimal upper bound on worst-case residual in the linear-fractional case are not known. We may, however, regularize this quantity using a method described in [29] for a related problem. For a given $\epsilon > 0$, define the bounded set

$$\mathcal{S}_\epsilon \triangleq \left\{ S \in \mathcal{S} \;\middle|\; \epsilon I \le S \le \frac{1}{\epsilon} I \right\},$$

where $\mathcal{S}$ is defined in (37). It is easy to show that restricting the condition number of variable S also bounds the variable G in the SDP (44). This yields the following result.

THEOREM 6.3. *An upper bound on the optimal worst-case residual can be obtained by computing the optimal value $\lambda(\epsilon)$ of the SDP*

$$\min_{S,G,\lambda} \lambda \text{ subject to } S \in \mathcal{S}_\epsilon,\ G \in \mathcal{G},\ (41). \tag{47}$$

The corresponding upper bound is a continuous function of $[A\ b]$. As $\epsilon \to 0$, the corresponding optimal value $\lambda(\epsilon)$ has a limit, equal to the optimal value of SDP (44).

As noted in Remark 5.1, the linear-fractional SRLS can be interpreted as a weighted LS and so can the above regularization method. Thus, the above method belongs to the class of Tikhonov (or weighted LS) regularization methods referred to in section 6.1, the weighting matrix being optimal for robustness.

7. Numerical examples. The following numerical examples were obtained using two different codes: for SDPs, we used the code **SP** [45], and a MATLAB interface to **SP** called **LMITOOL** [10]. For the (unstructured) RLS problems, we used the SOCP described in [28].

7.1. Complexity estimates of RLS. We first did "large-scale" experiments for the RLS problem in section 3. As mentioned in section 2.1, the number of iterations is almost independent of the size of the problem for SOCPs. We have solved problem (15) for uniformly generated random matrices A and vectors b with various sizes of n, m. Figure 1 shows the average number of iterations as well as the minimum and maximum number of iterations for various values of n, m. The experiments confirm the fact that the number of iterations is almost independent of problem size for the RLS problem.

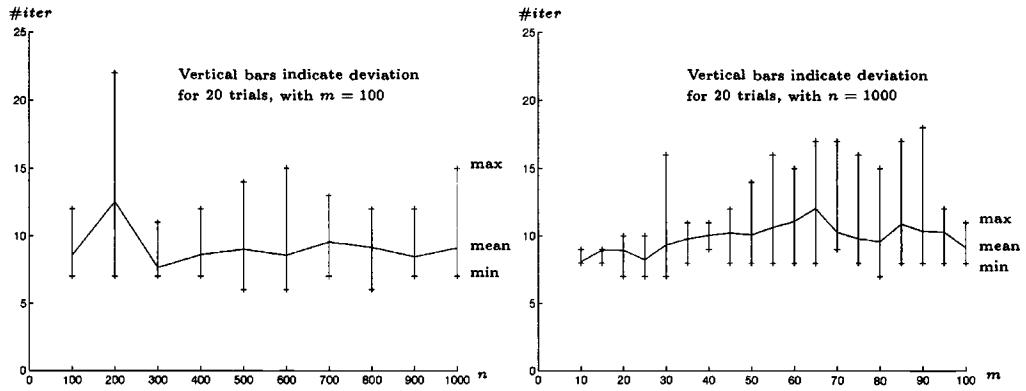


FIG. 1. *Average, minimum, and maximum number of iterations for various RLS problems using the SOCP formulation. In the left figure, we show these numbers for values of* n *ranging from* 100 *to* 1000. *For each value of* n, *the vertical bar indicates the minimum and maximum values obtained with* 20 *trials of* A, b, *with* $m = 100$. *In the right figure, we show these numbers for values of* m *ranging from* 11 *to* 100. *For each value of* n, *the vertical bar indicates the minimum and maximum values obtained with* 20 *trials of* A, b, *with* $n = 1000$. *For both plots, the plain curve is the mean value.*

7.2. LS, TLS, and RLS. We now compare the LS, TLS, and RLS solutions for

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix}^T, \quad b = \begin{bmatrix} 3 & 7 & 1 & 3 \end{bmatrix}^T.$$

On the left and right plots in Fig. 2, we show the four points (A_i, b_i) indicated with "+" signs, and the corresponding linear fits for LS problems (solid line), TLS problems (dotted line), and RLS problems for $\rho = 1, 2$ (dashed lines). The left plot gives the RLS solution with perturbations $[A + \Delta A, b + \Delta b]$, whereas the right plot considers perturbation in A only, $[A + \Delta A, b]$. In both plots, the worst-case points for the RLS solution are indicated by "o" for $\rho = 1$ and "$*$" for $\rho = 2$. As ρ increases, the slope of the RLS solution decreases and goes to zero when $\rho \to \infty$. The plot confirms Remark 3.3: the TLS solution is the most accurate and the least robust, and LS is intermediate.

In the case when we have perturbations in A only (right plot), we obtain an instance of a linear-fractional SRLS (with a full perturbation matrix), as mentioned in section 5.1. (It is also possible to solve this problem directly, as in section 3.) In this last case, of course, the worst-case perturbation can only move along the A-axis.

7.3. RLS and regularization. As mentioned in section 6, we may use RLS to regularize an ill-conditioned LS problem. Consider the RLS problem for

$$A = \begin{bmatrix} 3 & 1 & 4 \\ 0 & 1 & 1 \\ -2 & 5 & 3 \\ 1 & 4 & \alpha \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 2 \\ 1 \\ 3 \end{bmatrix}.$$

The matrix A is singular when $\alpha = 5$.

Figure 3 shows the regularizing effect of the RLS solution. The left (resp., right) figure shows the optimal worst-case residual (resp., norm of RLS solution) as a function of the parameter α for various values of ρ. When $\rho = 0$, we obtain the LS solution. The latter is not a continuous function of α, and both the solution norm

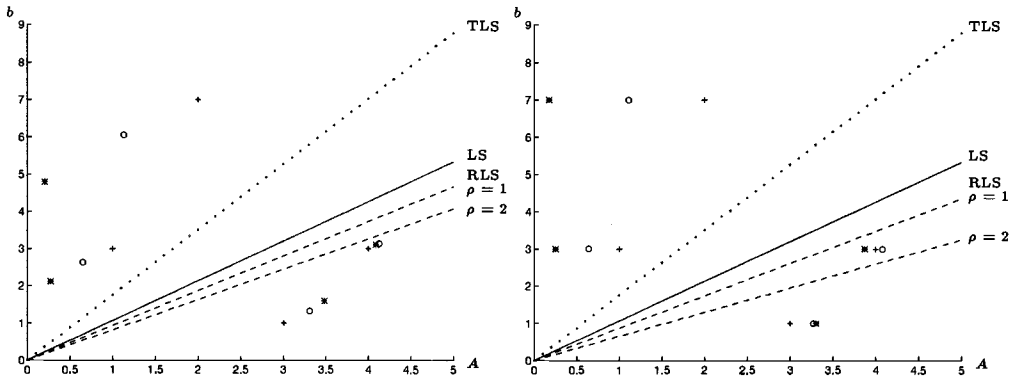


FIG. 2. *Least-squares (solid), total least-squares (dotted), and robust least-squares (dashed) solutions. The + signs correspond to the nominal [A b]. The left plot gives the RLS solution with perturbations* $[A+\Delta A, b+\Delta b]$*, whereas the right plot considers perturbation in A only,* $[A+\Delta A, b]$*. The worst-case perturbed points for the RLS solution are indicated by "o" for* $\rho = 1$ *and "∗" for* $\rho = 2$*.*

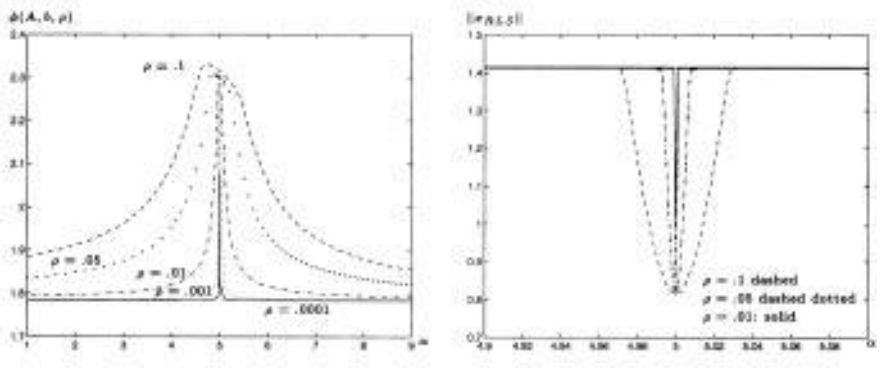


FIG. 3. *Optimal worst-case residual and norm of RLS solution versus* α *for various values of perturbation level* ρ*. For* $\rho = 0$ *(standard LS), the optimal residual and solution are discontinuous. The spike is smoothed as more robustness is asked for (that is, when* ρ *increases). On the right plot the curves for* $\rho = .001$ *and* $.0001$ *are not visible.*

and residual exhibit a spike for $\alpha = 5$ (when A becomes singular). For $\rho > 0$, the RLS solution is smooth. The spike is more and more flattened as ρ grows, which illustrates Theorem 6.1. For $\rho = \infty$, the optimal worst-case residual becomes flat (independent of α), and equal to $\|b\| + 1$, with $x_{\mathrm{RLS}} = 0$.

7.4. Robustness of LS solution. The next example illustrates that sometimes (precisely, if $b \in \mathbf{Range}(A)$) the LS solution is robust up to the perturbation level $\rho_{\min}$ defined in (22). This "natural" robustness of the LS solution degradates as the condition number of A grows. For $\varepsilon_A > 0$, consider the RLS problem for

$$A = \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon_A \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ .1 \end{bmatrix}.$$

We have considered six values of ε_A (which equals the inverse of the condition number of A) from .05 to .55. Table 1 shows the values of $\rho_{\min}$ (as defined in (22))

TABLE 1
Values of $\rho_{\min}$ for various ε_A.

curve #	1	2	3	4	5	6
ε_A	.05	.15	.25	.35	.45	.55
$\rho_{\min}$	0.06	0.34	0.78	1.12	1.28	1.35

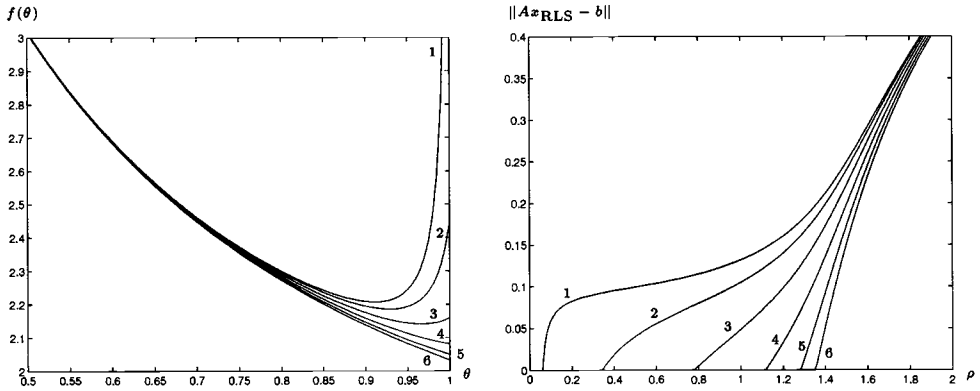


FIG. 4. *The left plot shows function $f(\theta)$ (as defined in* (20)*) for the six values of ε_A (for $\rho = 1$). The right plot gives the optimal RLS residuals versus ρ for the same values of ε_A. The labels* $1, \ldots, 6$ *correspond to values of ε_A given in Table* 1.

for the six values of ε_A. When the condition number of A grows, the robustness of the LS solution (measured by $\rho_{\min}$) decreases.

The right plot of Fig. 4 gives the worst-case residual versus the robustness parameter ρ for the six values of ε_A. The plot illustrates that for $\rho > \rho_{\min}$, the LS solution (in our case, $A^{-1}b$) differs from the RLS one. Indeed, for each curve, the residual remains equal to zero as long as $\rho \leq \rho_{\min}$. For example, the curve labeled "1" (corresponding to $\varepsilon_A = 0.05$) quits the x-axis for $\rho \geq \rho_{\min} = 0.06$.

The left plot of Fig. 4 corresponds to the RLS problem with $\rho = 1$ for various values of ε_A. This plot shows the various functions $f(\theta)$ as defined in (20). For each value of ε_A, the optimal θ (hence the RLS solution) is obtained by minimizing the function f. The three smallest values of ε_A induce functions f (as defined in (20)) that are minimal for $\theta < 1$. For the three others, the optimal θ is 1. This means that $\rho_{\min}$ is smaller than 1 in the first three cases and larger than 1 in the other cases. This is confirmed in Table 1.

7.5. Robust identification. Consider the following system identification problem. We seek to estimate the impulse response h of a discrete-time system from its input u and output y. Assuming that the system is single input and single output, linear, and of order m and that u is zero for negative time indices, y, u, and h are related by the convolution equations $Uh = y$, where

$$h = \begin{bmatrix} h(1) \\ \vdots \\ h(m) \end{bmatrix}, \quad y = \begin{bmatrix} y(1) \\ \vdots \\ y(m) \end{bmatrix}, \quad u = \begin{bmatrix} u(1) \\ \vdots \\ u(m) \end{bmatrix},$$

and U is a lower-triangular Toeplitz matrix whose first column is u. Assuming y, U are known exactly leads to a linear equation in h, which can be computed with standard

LS.

In practice, however, both y and u are subject to errors. We may assume, for instance, that the actual value of y is $y + \delta y$ and that of u is $u + \delta u$, where $\delta u, \delta y$ are unknown-but-bounded perturbations. For the perturbed matrices U, y write

$$\mathbf{U}(\delta) = U + \sum_{i=1}^{m} \delta u_i U_i, \quad \mathbf{y}(\delta) = y + \sum_{i=1}^{m} \delta y_i e_i,$$

where e_i, $i = 1, \ldots, m$ is the ith column of the $m \times m$ identity matrix and U_i are lower-triangular Toeplitz matrices with first column equal to e_i.

We first assume that the sum of the input and output energies is bounded, that is, $\|\delta\| \leq \rho$, where $\delta = [\delta u^T \ \delta y^T]^T \in \mathbf{R}^{2m}$, and $\rho \geq 0$ is given. We address the following SRLS problem:

$$\min_{h \in \mathbf{R}^m} \max_{\|\delta\| \leq \rho} \|\mathbf{U}(\delta)h - \mathbf{y}(\delta)\|. \tag{48}$$

As an example, we consider the following nominal values for y, u:

$$u = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}^T, \quad y = \begin{bmatrix} 4 & 5 & 6 \end{bmatrix}^T.$$

In Fig. 5, we have shown the optimal worst-case residual and that corresponding to the LS solution as given by solving problems (30) and (32), respectively. Since the LS solution has zero residual (U is invertible), we can prove (and check on the figure) that the worst-case residual grows linearly with ρ. In contrast, the RLS optimal worst-case residual has a finite limit as $\rho \to \infty$.

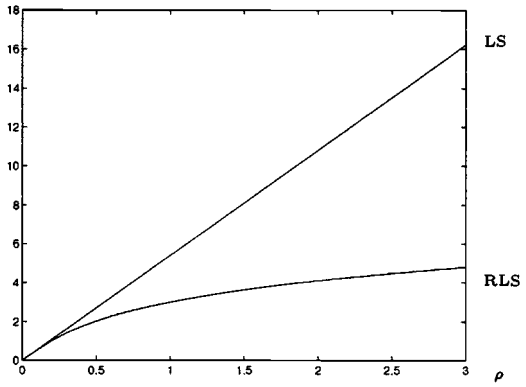


FIG. 5. *Worst-case residuals of LS and Euclidean-norm SRLS solutions for various values of perturbation level ρ. The worst-case residual for LS has been computed by solving problem* (30) *with $x = x_{\mathrm{LS}}$ fixed.*

We now assume that the perturbation bounds on y, u are not correlated. For instance, we consider problem (48), with the bound $\|\delta\| \leq \rho$ replaced with

$$\|\delta y\| \leq \rho, \quad \|\delta u\|_\infty \leq \rho.$$

Physically, the above bounds mean that the output energy and peak input are bounded.

This problem can be formulated as minimizing the worst-case residual (35), with

$$[A\ b] = \begin{bmatrix} 1 & 0 & 0 & 4 \\ 2 & 1 & 0 & 5 \\ 3 & 2 & 1 & 6 \end{bmatrix}, \quad L = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix},$$

$$R^T = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix},$$

and Δ has the following structure:

$$\Delta = \mathbf{diag}\left(\delta u_1 I_3, \delta u_2 I_2, \delta u_3, \begin{bmatrix} \delta y_1 & \times & \times \\ \delta y_2 & \times & \times \\ \delta y_3 & \times & \times \end{bmatrix}\right).$$

Here, the symbols $\times$ denote dummy elements of Δ that were added in order to work with a square perturbation matrix. The above structure corresponds to the set $\mathcal{D}$ in (36), with $s = [3\ 2\ 1]$.

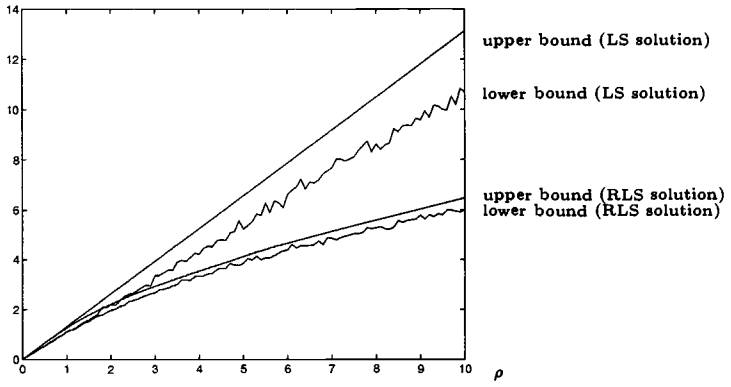


FIG. 6. *Upper and lower bounds on worst-case residuals for LS and RLS solutions. The upper bound for LS has been computed by solving the SDP* (38) *with* $x = x_{\mathrm{LS}}$ *fixed. The lower bounds correspond to the largest residuals* $\|\mathbf{U}(\delta^{\mathrm{trial}})x - \mathbf{y}(\delta^{\mathrm{trial}})\|$ *among* 100 *trial points* δ^{trial} *with* $x = x_{\mathrm{LS}}$ *and* $x = x_{\mathrm{RLS}}$.

In Fig. 6, we show the worst-case residual versus ρ, the uncertainty size. We show the curves corresponding to the values predicted by solving the SDP (43), with x variable (RLS solution), and x fixed to the LS solution x_{LS}. We also show lower bounds on the worst case, obtained using 100 trial points. This plot shows that, for the LS solution, our estimate of the worst-case residual is not exact, and the discrepancy grows linearly with uncertainty size. In contrast, for the RLS solution the estimate appears to be exact for every value of ρ.

7.6. Robust interpolation. The following example is a robust interpolation problem that can be formulated as a linear-fractional SRLS problem. For given integers $n \geq 1$, k, we seek a polynomial of degree $n-1$, $p(t) = x_1 + \cdots + x_n t^{n-1}$ that interpolates given points (a_i, b_i), $i = 1, \ldots, k$; that is,

$$p(a_i) = b_i, \quad i = 1, \ldots, k.$$

If we assume that (a_i, b_i) are known exactly, we obtain a linear equation in the unknown x, with a Vandermonde structure

$$\begin{bmatrix} 1 & a_1 & \dots & a_1^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & a_k & \dots & a_k^{n-1} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix},$$

which can be solved via standard LS.

Now assume that the interpolation points are not known exactly. For instance, we may assume that the b_i's are known, while the a_i's are parameter dependent:

$$a_i(\delta) = a_i + \delta_i, \quad i = 1, \dots, k,$$

where the δ_i's are unknown but bounded, $|\delta_i| \leq \rho$, $i = 1, \dots, k$, where $\rho \geq 0$ is given. We seek a robust interpolant, that is, a solution x that minimizes

$$\max_{\|\delta\|_\infty \leq \rho} \|\mathbf{A}(\delta)x - b\|,$$

where

$$\mathbf{A}(\delta) = \begin{bmatrix} 1 & a_1(\delta) & \dots & a_1(\delta)^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & a_k(\delta) & \dots & a_k(\delta)^{n-1} \end{bmatrix}.$$

The above problem is a linear-fractional SRLS problem. Indeed, it can be shown that

$$\begin{bmatrix} \mathbf{A}(\delta) & b \end{bmatrix} = \begin{bmatrix} \mathbf{A}(0) & b \end{bmatrix} + L\Delta(I - D\Delta)^{-1} \begin{bmatrix} R_A & 0 \end{bmatrix},$$

where

$$L = \bigoplus_{i=1}^{k} \begin{bmatrix} 1 & a_i & \dots & a_i^{n-2} \end{bmatrix}, R_A = \begin{bmatrix} R_1 \\ \vdots \\ R_k \end{bmatrix}, \quad D = \bigoplus_{i=1}^{k} D_i, \quad \Delta = \bigoplus_{i=1}^{k} \delta_i I_{n-1},$$

and, for each i, $i = 1, \dots, k$,

$$R_i = \begin{bmatrix} 0 & 1 & a_i & \dots & a_i^{n-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & a_i \\ 0 & \dots & \dots & 0 & 1 \end{bmatrix} \in \mathbf{R}^{(n-1)\times n},$$

$$D_i = \begin{bmatrix} 0 & 1 & a_i & \dots & a_i^{n-3} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & a_i \\ \vdots & & & \ddots & 1 \\ 0 & \dots & \dots & \dots & 0 \end{bmatrix} \in \mathbf{R}^{(n-1)\times(n-1)}.$$

(Note that $\det(I - D\Delta) \neq 0$, since D is strictly upper triangular.)

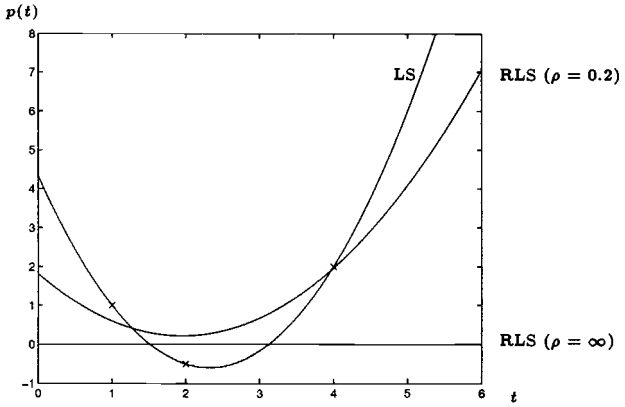


FIG. 7. *Interpolation polynomials: LS and RLS solutions for* $\rho = 0.2$. *The LS solution interpolates the points exactly, while the RLS one guarantees a worst-case residual error less than* 1.1573. *For* $\rho = \infty$, *the RLS solution is the zero polynomial.*

In Fig. 7, we have shown the result $n = 3$, $k = 1$, and

$$a_1 = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}, \quad b_1 = \begin{bmatrix} 1 \\ -0.5 \\ 2 \end{bmatrix}, \quad \rho = 0.2.$$

The LS solution is very accurate (zero nominal residual: every point is interpolated exactly) but has a (predicted) worst-case residual of 1.7977. The RLS solution trades off this accuracy (only one point interpolated and nominal residual of 0.8233) for robustness (with a worst-case residual less than 1.1573). As $\rho \to \infty$, the RLS interpolation polynomial becomes more and more horizontal. (This is consistent with the fact that we allow perturbations on vector a only.) In the limit, the interpolation polynomial is the solid line $p(t) = 0$.

8. Conclusions. This paper shows that several RLS problems with unknown-but-bounded data matrices are amenable to (convex) SOCP or SDP. The implication is that these RLS problems can be solved in polynomial time and efficiently in practice.

When the perturbation enters linearly in the data matrices, and its size is measured by Euclidean norm, or in a linear-fractional problem with full perturbation matrix Δ, the method yields the *exact* value of the optimal worst-case residual. In the other cases we have examined (such as arbitrary rational dependence of data matrices on the perturbation parameters), computing the worst-case residual is NP-complete. We have shown how to compute and optimize, using SDP, an upper bound on the worst-case residual that takes into account structure information.

In the unstructured case, we have shown that both the worst-case residual and the (unique) RLS solution are continuous. The unstructured RLS can be interpreted as a regularization method for ill-conditioned problems. A striking fact is that the cost of the RLS solution is equal to a small number of least-squares problems arising in classical Tikhonov regularization approaches. This method provides a rigorous way to compute the optimal parameter from the data and associated perturbation bounds. Similar (weighted) least-squares interpretations and continuity results were given for the structured case.

In our examples, we have demonstrated the use of an SOCP code [27] and a general-purpose semidefinite programming code **SP** [45]. Future work could be devoted to writing special code that exploits the structure of these problems in order to further increase the efficiency of the method. For instance, it seems that in many problems the perturbation matrices are sparse and/or have special (e.g., Toeplitz) structure.

The method can be used for several related problems.

- *Constrained RLS.* We may consider problems where additional (convex) constraints are added on the vector x. (Such constraints arise naturally in, e.g., image processing.) For instance, we may consider problem (1) with an additional linear (resp., quadratic convex) constraint $(Cx)_i \geq 0$, $i = 1, \ldots, q$ (resp., $x^T Q x \leq 1$), where C (resp., $Q \geq 0$) is given. To solve such a problem, it suffices to add the related constraint to corresponding SOCP or SDP formulation. (Note that the SVD approach of section 3.3 fails in this case.)
- *RLS problems with other norms.* We may consider RLS problems in which the worst-case residual errors measured in other norms such as the maximum (l_∞) norm.
- *Matrix RLS.* We may, of course, derive similar results when the constant term b is a matrix. The worst-case error can be evaluated in a variety of norms.
- *Error-in-variables RLS.* We may consider problems where the solution x is also subject to uncertainty (due to implementation and/or quantization errors). That is, we may consider a worst-case residual of the form

$$\max_{\|\Delta x\| \leq \rho_1} \max_{\|\Delta A \ \Delta b\|_F \leq \rho_2} \|(A + \Delta A)(x + \Delta x) - (b + \Delta b)\|,$$

  where ρ_i, $i = 1, 2$, are given. We may compute (and optimize) upper bounds on the above quantity using SDP. This subject is examined in [25].

Appendix A. Proof of Theorem 4.1. Introduce the eigendecomposition of F and a related decomposition for g:

$$F = \tau I - U \begin{bmatrix} \tau - \lambda_{\max}(F) & 0 \\ 0 & \tau I - \Sigma \end{bmatrix} U^T, \quad U^T g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix},$$

where $\tau > \|\Sigma\|$, $\Sigma \in \mathbf{R}^{r \times r}$, $\Sigma > 0$, and $g_2 \in \mathbf{R}^r$. When $\tau > \lambda_{\max}(F)$, inequality (29) writes

$$\lambda \geq h + \tau + \frac{g_1^T g_1}{\tau - \lambda_{\max}(F)} + g_2^T (\tau I - \Sigma)^{-1} g_2. \tag{A.49}$$

If $\tau = \lambda_{\max}(F)$ at the optimum, then $g_1 = 0$, and there exists a nonzero vector u such that $(\tau I - F)u = 0$. From inequality (29), we conclude that $g^T u = 0$. In other words, $\lambda_{\max}(F)$ is not (F, g)-controllable, and u is an eigenvector that proves this uncontrollability. Using $g_1 = 0$ in (A.49), we obtain the optimal value of λ in this case:

$$\lambda = h + \tau + g_2^T (\tau I - \Sigma)^{-1} g_2.$$

Thus, the worst-case residual can be computed as claimed in the theorem.

For every pair (λ, τ) that is optimal for problem (29), we can compute a worst-case perturbation as follows. Define

$$\delta_0 = (\tau I - F)^\dagger g.$$

We have $\tau > \lambda_{\max}(F)$ at the optimum if and only if $\lambda_{\max}(F)$ is (F, g)-controllable (that is, $g_2 \neq 0$) or if $\lambda_{\max}(F)$ is not (F, g)-controllable and the function f defined in (31) satisfies

$$\frac{df}{d\tau}(\lambda_{\max}(F)) = 1 - g^T(\lambda_{\max}(F)I - F)^{2\dagger}g < 0.$$

In this case, the optimal τ satisfies

$$1 = g^T(\tau I - F)^{-2}g; \tag{A.50}$$

that is, $\|\delta_0\| = 1$. Using this and (A.50), we obtain

$$\begin{bmatrix} 1 \\ \delta_0 \end{bmatrix}^T \begin{bmatrix} h & g^T \\ g & F \end{bmatrix} \begin{bmatrix} 1 \\ \delta_0 \end{bmatrix} = \lambda.$$

This proves that δ_0 is a worst-case perturbation.

If $\tau = \lambda_{\max}(F)$ at the optimum, then

$$\frac{df}{d\tau}(\lambda_{\max}(F)) = 1 - g^T(\lambda_{\max}(F)I - F)^{2\dagger}g \geq 0,$$

which implies that $\|\delta_0\| \leq 1$. Since $\tau = \lambda_{\max}(F)$, there exists a vector u such that $(\tau I - F)u = 0$, $g^T u = 0$. Without loss of generality, we may assume that the vector $\delta = \delta_0 + u$ satisfies $\|\delta\| = 1$. We have

$$\begin{aligned} &\begin{bmatrix} 1 \\ \delta \end{bmatrix}^T \begin{bmatrix} h & g^T \\ g & F \end{bmatrix} \begin{bmatrix} 1 \\ \delta \end{bmatrix} = \tau\delta^T\delta - \delta^T(\tau I - F)\delta + 2\delta_0^T g + h \\ &= h + \tau + g^T(\tau I - F)^{\dagger}g - 2u^T(\tau I - F)\delta_0 - u^T(\tau I - F)u = \lambda. \end{aligned}$$

This proves that δ defined above is a worst-case perturbation.

In both cases seen above (τ equals $\lambda_{\max}(F)$ or not), a worst-case perturbation is any vector δ such that

$$(\tau I - F)\delta = g, \quad \|\delta\| = 1.$$

(We have just shown that the above equations always have a solution δ when τ is optimal.) This ends our proof. □

Appendix B. Proof of Lemma 5.1. We use the following result, due to Nemirovsky [32].

LEMMA B.1. *Let $\Gamma(p, a)$ be a scalar function of positive integer p and p-dimensional vector* a *such that, first, Γ is well defined and takes rational values from $(0, \|a\|^{-2})$ for all positive integers p and all p-dimensional vectors a with $\|a\| \leq 0.1$ and, second, the value of this function at a given pair (p, a) can be computed in time polynomial in p and the length of the standard representation of the (rational) vector a. Then the problem $\mathcal{P}_\Gamma(p, a)$: given an integer $p \geq 0$ and $a \in \mathbf{R}^p$, $\|a\| \leq 0.1$, with rational positive entries, determine whether*

$$p \leq \max_{\|\delta\|_\infty \leq 1} \delta^T(I - \Gamma(p, a)aa^T)\delta \tag{B.51}$$

is NP-complete. Besides this, either (B.51) *holds, or*

$$p - \frac{\Gamma(p, a)}{d(a)^2} \geq \max_{\|\delta\|_\infty \leq 1} \delta^T(I - aa^T)\delta,$$

where $d(a)$ is the smallest common denominator of the entries of a.

To prove our result, it suffices to show that for some appropriate function Γ satisfying the conditions of Lemma B.1, for any given p, a, we can reduce the problem $\mathcal{P}_\Gamma(p, a)$ to problem $\mathcal{P}(\mathbf{A}, \mathbf{b}, \mathcal{D}, x)$ in polynomial time. Set

$$\Gamma(p, a) = \frac{2a^T a + 1}{(a^T a + 1)^2}.$$

This function satisfies all requirements of Lemma B.1, so problem $\mathcal{P}_\Gamma(p, a)$ is NP-hard.

Given p, a, $\|a\| \leq 0.1$ with rational positive entries, set $\mathbf{A}$, $\mathbf{b}$, $\mathcal{D}$ and x as follows. First, set $\mathcal{D}$ to be the set of diagonal matrices of $\mathbf{R}^{p\times p}$. Set $A = 0$, $b = 0$, $R_A = 0$, $R_b = [1 \dots 1]^T$, $D = 0$, $x = 0$, and

$$L = I - \frac{aa^T}{1 + a^T a}.$$

Finally, set $\mathbf{A}$, $\mathbf{b}$ as in (34) and $\lambda = p - \Gamma(p, a)/d(a)^2$. When $\rho = 1$, the worst-case residual for this problem is

$$r_{\mathcal{D}}(\mathbf{A}, \mathbf{b}, 1, x)^2 = \max_{\|\delta\|_\infty \leq 1} \|L\delta\|^2 = \max_{\|\delta\|_\infty \leq 1} \delta^T (I - \Gamma(p, a)aa^T)\delta.$$

Our proof is now complete.

Appendix C. Proof of Theorem 5.3. In this section, we only prove Theorem 5.3. The proof of Theorem 5.2 follows the same lines. We start from problem (43), the dual of which is the maximization of $2(b^T w + R_b^T u)$ subject to

$$\mathcal{Z} = \begin{bmatrix} Z & Y & w \\ Y^T & V & u \\ w^T & u^T & t \end{bmatrix} \geq 0 \tag{C.52}$$

and the linear constraints

$$\mathbf{Tr} Z = 1 - t, \tag{C.53}$$

$$\forall S \in \mathcal{S}, \ \ \mathbf{Tr} S(V - L^T Z L - D^T Y^T L - L^T Y D - D^T V D) = 0, \tag{C.54}$$

$$A^T w + R_A^T u = 0, \tag{C.55}$$

$$\forall G \in \mathcal{G}, \ \ \mathbf{Tr} G(YL - L^T Y^T - D^T V + V D) = 0. \tag{C.56}$$

Since both primal and dual problems are strictly feasible, all primal and dual feasible points are optimal if and only if $\mathcal{Z}\mathcal{F}(\lambda, S, G, x) = 0$, where $\mathcal{F}$ is defined in (38) (see [46]). One obtains, in particular,

$$Jw + t(Ax - b) - L\Gamma u = 0, \tag{C.57}$$

$$(Ax - b)^T w + t\lambda + z^T R^T u = 0, \tag{C.58}$$

$$-\Gamma^T L^T w + Rz + \Sigma u = 0, \tag{C.59}$$

where $z = [x^T \ -1]^T$, $J = \lambda I - LSL^T$, $\Sigma = S + DG - GD^T - DSD^T$, and $\Gamma = SD^T - G$.

Using equation (C.58) and (C.55), we obtain

$$t\lambda = -(Ax - b)^T w - z^T R^T u = b^T w + R_b^T u, \tag{C.60}$$

which implies that $t = 1/2$ from equality of the primal and dual objectives (the trivial case $\lambda = 0$ can be easily ruled out).

Assume that the matrix Θ defined in (39) is positive definite at the optimum. From equations (C.57)–(C.59), we deduce that the dual variable $\mathcal{Z}$ is rank one:

$$\mathcal{Z} = 2vv^T \text{ with } v = \left[\begin{array}{cccc} w & u & 1/2 \end{array} \right]^T. \tag{C.61}$$

Using (C.57) and (C.59), we obtain

$$\Theta \left[\begin{array}{c} w \\ u \end{array} \right] = \frac{1}{2} \left[\begin{array}{c} Ax - b \\ R_A x - R_b \end{array} \right].$$

From (C.55), it is easy to derive the expression (42) for the optimal x in the case when $\Theta > 0$ at the optimum and R_A is full rank.

We now show that the upper bound is exact at the optimum in this case. If we use condition (C.54) and the expression for Z, V deduced from (C.53), we obtain

$$u^T S u = (L^T w + D^T u)^T S (L^T w + D^T u) \text{ for every } S \in \mathcal{S}.$$

This implies that there exists $\Delta \in \mathcal{D}$, $\Delta^T \Delta = I$, such that $u = \Delta^T (L^T w + D^T u)$. Since $\Theta > 0$, a straightforward application of Lemma 2.3 shows that $\det(I - D\Delta) \neq 0$, so we obtain

$$u^T = w^T L \Delta (I - D\Delta)^{-1}.$$

Define $M = [A\ b]$ and recall $z = [x^T\ -1]^T$. Since $Z = 2ww^T$ (from (C.61)) and $\mathbf{Tr} Z = 1 - t = 1/2$ (from (C.53)), we have $\|w\| = 1/2$. We can now compute

$$\begin{aligned} w^T (M + L\Delta(I - D\Delta)^{-1} R) z &= w^T (Ax - b) + w^T L \Delta (I - D\Delta)^{-1} R z \\ &= w^T (Ax - b) + u^T R z \\ &= -\frac{\lambda}{2} \text{ (from (C.55) and (C.60)).} \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{\lambda}{2} = \left| w^T (M + L\Delta(I - D\Delta)^{-1} R) z \right| &\leq \|w\| \left\| (M + L\Delta(I - D\Delta)^{-1} R) z \right\| \\ &\leq \|w\| \lambda \text{ (since } \Delta \in \mathcal{D}, \|\Delta\| \leq 1) \\ &= \frac{\lambda}{2} \left(\text{from} \|w\| = \frac{1}{2} \right). \end{aligned}$$

We obtain $\lambda = \left\| (M + L\Delta(I - D\Delta)^{-1} R) z \right\|$, which proves that the matrix Δ is a worst-case perturbation.

Acknowledgments. The authors wish to thank the anonymous reviewers for their precious comments, which led to many improvements over the first version of this paper. We are particularly indebted to the reviewer who pointed out the SOCP formulation for the unstructured problem. We also thank G. Golub and R. Tempo for providing us with some related references and A. Sayed for sending us the preliminary draft [5]. The paper has also benefited from many fruitful discussions with S. Boyd, F. Oustry, B. Rottembourg, and L. Vandenberghe.

REFERENCES

[1] K. D. Andersen, *An efficient Newton barrier method for minimizing a sum of Euclidean norms*, SIAM J. Optim., 6 (1996), pp. 74–95.

[2] Å. Björck, *Component-wise perturbation analysis and error bounds for linear least squares solutions*, BIT, 31 (1991), pp. 238–244.

[3] J. F. Bonnans, R. Cominetti, and A. Shapiro, *Sensitivity analysis of optimization problems under abstract constraints*, SIAM J. Optim., submitted.

[4] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, in Studies in Applied Mathematics, SIAM, Philadelphia, PA, 1994.

[5] S. Chandrasekaran, G. H. Golub, M. Gu, and A. H. Sayed, *A new linear least-squares type model for parameter estimation in the presence of data uncertainties*, SIAM J. Matrix Anal. Appl., submitted.

[6] G. E. Coxson and C. L. DeMarco, *Computing the Real Structured Singular Value is NP-Hard*, Tech. report ECE-92-4, Dept. of Elec. and Comp. Eng., University of Wisconsin-Madison, Madison, WI, June 1992.

[7] B. de Moor, *Structured total least squares and* $\mathbf{L}_2$ *approximation problems*, Linear Algebra Appl., 188–189 (1993), pp. 163–207.

[8] G. Demoment, *Image reconstruction and restoration: Overview of common estimation problems*, IEEE Trans. Acoustic Speech and Signal Processing, 37 (1989), pp. 2024–2036.

[9] J. Doyle, M. Newlin, F. Paganini, and J. Tierno, *Unifying robustness analysis and system ID*, in Proc. IEEE Conf. on Decision and Control, December 1994, pp. 3667–3672.

[10] L. El Ghaoui, R. Nikoukhah, and F. Delebecque, `LMITOOL:` *A Front-End for LMI Optimization, User's Guide*, February 1995. Available via anonymous ftp from ftp.ensta.fr/pub/elghaoui/lmitool.

[11] L. Elden, *Algorithms for the regularization of ill conditioned least-squares problems*, BIT, 17 (1977), pp. 134–145.

[12] L. Elden, *Perturbation theory for the least-squares problem with linear equality constraints*, BIT, 24 (1985), pp. 472–476.

[13] M. K. H. Fan, A. L. Tits, and J. C. Doyle, *Robustness in the presence of mixed parametric uncertainty and unmodeled dynamics*, IEEE Trans. Automat. Control, 36 (1991), pp. 25–38.

[14] R. D. Fierro and J. R. Bunch, *Collinearity and total least squares*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1167–1181.

[15] M. Furuya, H. Ohmori, and A. Sano, *Optimization of weighting constant for regularization in least squares system identification*, Trans. Inst. Elec. Inform. Comm. Eng. A, J72A (1989), pp. 1012–1015.

[16] M. X. Goemans and D. P. Williamson, .878-*approximation for MAX CUT and MAX* 2*SAT*, in Proc. 26th ACM Symp. Theor. Computing, 1994, pp. 422–431.

[17] 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.

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

[19] G. H. Golub and U. von Matt, *Quadratically constrained least squares and quadratic problems*, Numer. Math., 59 (1991), pp. 561–580.

[20] M. L. Hambaba, *The robust generalized least-squares estimator*, Signal Processing, 26 (1992), pp. 359–368.

[21] M. Hanke and P. C. Hansen, *Regularization methods for large-scale problems*, Surveys on Mathematics for Industry, 3 (1993), pp. 253–315.

[22] D. J. Higham and N. J. Higham, *Backward error and condition of structured linear systems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 162–175.

[23] B. R. Hunt, *The application of constrained least-squares estimation to image restoration by digital computer*, IEEE Trans Comput., C-22 (1973), pp. 805–812.

[24] T. Iwasaki and R. E. Skelton, *All controllers for the general* $\mathbf{H}_\infty$ *control problem: LMI existence conditions and state space formulas*, Automatica, 30 (1994), pp. 1307–1317.

[25] C. Jacquemont, *Error-in-Variables Robust Least-Squares*, Tech. rep., Ecole Nat. Sup. Techniques Avancies, 32, Bd. Victor, 75739 Paris, France, December 1995.

[26] G. Lady and J. Maybee, *Qualitatively invertible matrices*, J. Math. Social Sciences, 6 (1983), pp. 397–407.

[27] H. Lebret, *Synthèse de diagrammes de réseaux d'antennes par optimisation convexe*, Ph.D. thesis, UFR Structure et Propriétés de la Matière, mention Electronique, Université de

Rennes I, France, 1994.
[28] H. Lebret, *Antenna pattern synthesis through convex optimization*, in Advanced Signal Processing Algorithms, Proc. SPIE 2563, F. T. Luk, ed., 1995, pp. 182–192.
[29] L. Lee and A. Tits, *On continuity/discontinuity in robustness indicators*, IEEE Trans. Automat. Control, 38 (1993), pp. 1551–1553.
[30] K. Miller, *Least squares methods for ill-posed problems with a prescribed bound*, SIAM J. Math. Anal., 1 (1970), pp. 52–74.
[31] M. Z. Nashed, *Operator-theoretic and computational approaches to ill-posed problems with applications to antenna theory*, IEEE Trans. Antennas and Propagation, 29 (1981), pp. 220–231.
[32] A. Nemirovsky, *Several NP-hard problems arising in robust stability analysis*, Mathematics of Control, Signals, and Systems, 6 (1993), pp. 99–105.
[33] Y. Nesterov and A. Nemirovsky, *Interior Point Polynomial Methods in Convex Programming: Theory and Applications*, SIAM, Philadelphia, PA, 1994.
[34] J. Norton, *Identification and application of bounded parameter models*, Automatica, 31 (1987), pp. 497–507.
[35] F. Oustry, L. El Ghaoui, and H. Lebret, *Robust solutions to uncertain semidefinite programs*, SIAM J. Optim., submitted, 1996.
[36] C. Papadimitriou and M. Yannakakis, *Optimization, approximation and complexity classes*, J. Comput. System Sci., 43 (1991), pp. 425–440.
[37] J. R. Partington and P. M. Mäkilä, *Worst-case analysis of the least-squares method and related identification methods*, Systems Control Lett., 24 (1995), pp. 193–200.
[38] S. Poljak and J. Rohn, *Checking robust nonsingularity is NP-hard*, Math. Control Signals Systems, 6 (1993), pp. 1–9.
[39] B. L. Shader, *Least squares sign-solvability*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1056–1073.
[40] R. Smith and J. Doyle, *Model validation: A connection between robust control and identification*, IEEE Trans. Automat. Control, 37 (1992), pp. 942–952.
[41] R. J. Stern and H. Wolkowicz, *Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations*, SIAM J. Optim., 5 (1995), pp. 286–313.
[42] R. Tempo, *Worst-case optimality of smoothing algorithms for parametric system identification*, Automatica, 31 (1995), pp. 759–764.
[43] A. Tikhonov and V. Arsenin, *Solutions of Ill-Posed Problems*, Wiley, New York, 1977.
[44] S. Van Huffel and J. Vandewalle, *The total least squares problem: Computational aspects and analysis*, in Frontiers in Applied Mathetics 9, SIAM, Philadelphia, PA, 1991.
[45] L. Vandenberghe and S. Boyd, **SP,** *Software for Semidefinite Programming, User's Guide*, Dec. 1994. Available via anonymous ftp from isl.stanford.edu under /pub/boyd/semidef_prog.
[46] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.
[47] M. E. Zervakis and T. M. Kwon, *Robust estimation techniques in regularized image restoration*, Op. Eng., 31 (1992), pp. 2174–2190.
[48] K. Zhou, J. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice–Hall, Englewood Cliffs, NJ, 1995.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 4, pp. 1065–1081, October 1997

016

LOCALITY OF REFERENCE IN LU DECOMPOSITION WITH PARTIAL PIVOTING *

SIVAN TOLEDO†

Abstract. This paper presents a new partitioned algorithm for LU decomposition with partial pivoting. The new algorithm, called the recursively partitioned algorithm, is based on a recursive partitioning of the matrix. The paper analyzes the locality of reference in the new algorithm and the locality of reference in a known and widely used partitioned algorithm for LU decomposition called the right-looking algorithm. The analysis reveals that the new algorithm performs a factor of $\Theta(\sqrt{M/n})$ fewer I/O operations (or cache misses) than the right-looking algorithm, where n is the order of the matrix and M is the size of primary memory. The analysis also determines the optimal block size for the right-looking algorithm. Experimental comparisons between the new algorithm and the right-looking algorithm show that an implementation of the new algorithm outperforms a similarly coded right-looking algorithm on six different RISC architectures, that the new algorithm performs fewer cache misses than any other algorithm tested, and that it benefits more from Strassen's matrix-multiplication algorithm.

Key words. LU factorization, Gaussian elimination, partial pivoting, locality of reference, cache misses

AMS subject classifications. 15A23, 65F05, 65Y10, 65Y20

PII. S0895479896297744

1. Introduction.

Algorithms that partition dense matrices into blocks and operate on entire blocks as much as possible are key to obtaining high performance on computers with hierarchical memory systems. Partitioning a matrix into blocks creates temporal locality of reference in the algorithm and reduces the number of words that must be transferred between primary and secondary memories. This paper describes a new partitioned algorithm for LU factorization with partial pivoting, called the *recursively partitioned* algorithm. The paper also analyzes the number of data transfers in a popular partitioned LU-factorization algorithm, the so-called *right-looking* algorithm, which is used in LAPACK [1]. The performance characteristics of other popular partitioned LU-factorization algorithms, in particular Crout and the left-looking algorithm used in the NAG library [4], are similar to those of the right-looking algorithm so they are not analyzed.

The analysis of the two algorithms leads to two interesting conclusions. First, there is a simple system-independent formula for choosing the block size for the right-looking algorithm which is almost always optimal. Second, the recursively partitioned algorithm generates asymptotically less memory traffic between memories than the right-looking algorithm, even if the block size for the right-looking algorithm is chosen optimally. Numerical experiments indicate that the recursively partitioned algorithm generates fewer cache misses and runs faster than the right-looking algorithm.

The recursively partitioned algorithm computes the LU decomposition with partial pivoting of an n-by-m matrix while transferring only $\Theta(nm^2/\sqrt{M} + nm \lg m)$

*Received by the editors January 26, 1996; accepted for publication (in revised form) by A. Edelman November 12, 1996. Parts of this research were performed while the author was a postdoctoral fellow at the IBM T. J. Watson Research Center and a postdoctoral associate at the MIT Laboratory for Computer Science. The work at MIT was supported in part by ARPA grant N00014-94-1-0985.
http://www.siam.org/journals/simax/18-4/29774.html

†Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304 (toledo@parc.xerox.com).

words between primary and secondary memories, where M is the size of the primary memory. The right-looking algorithm, on the other hand, transfers at least $\Theta(\max(nm^2/\sqrt{M}, nm^{1.5}))$ words. The number of words actually transferred by conventional algorithms depends on a parameter r, which is not chosen optimally in LAPACK. The new algorithm is optimal in the sense that the number of words that it transfers is asymptotically the same as the number transferred by partitioned (or blocked) algorithms for matrix multiplication and solution of triangular systems (at least when the number of columns is not very small compared with the size of primary memory). The right-looking algorithm achieves such performance only when the matrix is so large that a few rows fill the primary memory.

The recursively partitioned algorithm has other advantages over conventional algorithms. It has no block-size parameter that must be tuned in order to achieve high performance. Since it is recursive, it is likely to perform better when the memory system has more than two levels, for example, on computer systems with two levels of cache or with both cache and virtual memory.

To understand the main idea behind the new algorithm, let us look first at the conventional right-looking LU-factorization algorithm. The algorithm decomposes the input matrix into $\lceil n/r \rceil$ blocks of at most r columns. Starting from the leftmost block of columns, the algorithm iteratively factors a block of r columns using a column-oriented algorithm. After a block is factored, the algorithm updates the entire trailing submatrix. The parameter r must be carefully chosen to minimize the number of words transferred between memories. If r is larger than M/n, many words must be transferred when a block of columns is factored. If r is too small, many trailing submatrices must be updated, and most of the updates require the entire trailing submatrix to be read from secondary memory.

The main insight behind the recursively partitioned algorithm is that there is no need to update the entire trailing submatrix after a block of columns is factored. After factoring the first column of the matrix, the algorithm updates just the next column to the right, which enables it to proceed. Once the second column is factored, we must apply the updates from the first two columns before we can proceed. The algorithm updates two more columns and proceeds. Once four columns are factored, they are used to update four more, and so on. In other words, the algorithm does not look all the way to the right every time a few columns are factored. As we shall see below, this short-sighted approach pays off.

From another point of view, the new algorithm is a recursive algorithm. We know that the larger r (the number of columns in a block), the smaller the number of data transfers required for updating trailing submatrices. The algorithm therefore chooses the largest possible size, $r = m/2$. If that many columns do not fit within primary memory, they are factored recursively using the same algorithm, rather than being factored using a naive column-oriented algorithm. Once the left $m/2$ columns are factored, they are used to update the right $m/2$ columns which are subsequently factored.

The rest of the paper is organized as follows. Section 2 describes and analyzes the recursively partitioned algorithm. Section 3 analyzes the block-column right-looking algorithm. The actual performance of LAPACK's right-looking algorithm and the performance of the recursively partitioned algorithm are compared in section 4 on several high-end workstations. Section 5 concludes the paper with a discussion of the results and of related research.

2. Recursively partitioned LU factorization. The recursively partitioned algorithm is not only more efficient than conventional partitioned algorithms, but it is also simpler to describe and analyze. This section first describes the algorithm, and then analyzes the complexity of the algorithm in terms of arithmetic operations and in terms of the amount of data transferred between memories during its execution.

The algorithm. The algorithm factors an n-by-m matrix A into an n-by-n permutation matrix P, an n-by-m unit lower triangular matrix L (that is, L's upper triangle is all zeros), and an m-by-m upper triangular matrix U, such that $PA = LU$. A is treated as a block matrix

$$A = \begin{bmatrix} A_{11} & A_{21} \\ A_{21} & A_{22} \end{bmatrix},$$

where A_{11} is a square matrix of order $m/2$-by-$m/2$.

1. If $m = 1$ then factor (that is, perform pivoting and scaling)

$$P_1 \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} = \begin{bmatrix} L_{11} \\ L_{21} \end{bmatrix} U_{11}$$

   and return.
2. Else, recursively factor

$$P_1 \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} = \begin{bmatrix} L_{11} \\ L_{21} \end{bmatrix} U_{11}.$$

3. Permute

$$\begin{bmatrix} A'_{12} \\ A'_{22} \end{bmatrix} \leftarrow P_1 \begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix}.$$

4. Solve the triangular system $L_{11}U_{12} = A'_{12}$ for U_{12}.
5. $A''_{22} \leftarrow A'_{22} - L_{21}U_{12}$.
6. Recursively factor $P_2A''_{22} = L_{22}U_{22}$.
7. Permute $L'_{21} \leftarrow P_2L_{21}$.
8. Return

$$P_2P_1 \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L'_{21} & L_{22} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix}.$$

Complexity analysis. It is not hard to see that the algorithm is numerically equivalent to the conventional column-oriented algorithm. Therefore, the algorithm has the same numerical properties as the conventional algorithm, and it performs same number of floating point operations, about $nm^2 - m^3/3$. In fact, all the variants of the LU-factorization algorithm discussed in this paper are essentially different schedules for the same algorithm. That is, they all have the same data-flow graph.

We now analyze the number of words that must be transferred between the primary and secondary memories for $n \geq m$. The size of primary memory is denoted by M. For ease of exposition, we assume that the number of columns is a power of two. We denote the number of words that the algorithm must transfer between memories by $\mathbf{IO_{RP}}(n, m)$. We denote the number of words that must be transferred to solve an n-by-n triangular linear system with m right-hand sides where the solution overwrites the right-hand side by $\mathbf{IO_{TS}}(n, m)$. We denote the number of words that must be

transferred to multiply an n-by-m matrix by an m-by-k matrix and add the result to an n-by-k matrix by $\mathbf{IO_{MM}}(n, m, k)$.

Since the factorization algorithm uses matrix multiplication and solution of triangular linear system as subroutines, the number of I/Os it performs depends on the number of I/Os performed by these subroutines. A partitioned algorithm for solving triangular linear systems performs at most

$$\mathbf{IO_{TS}}(m,m) \leq \begin{cases} 2.5m^2 & \text{if } m \leq \sqrt{M/3}, \\ \frac{m^3}{\sqrt{M/3}} + m^2 & \text{if } m \geq \sqrt{M/3} \end{cases} \tag{2.1}$$

I/Os. The actual number of I/Os performed is smaller, since the real crossover point is $\sqrt{M/2}$, not $\sqrt{M/3}$. Incorporating the improved bound into the analysis complicates the analysis with little effect on the final outcome. The number of I/Os performed by a standard matrix-multiplication algorithm is at most

$$\mathbf{IO_{MM}}(n,n,m) \leq \begin{cases} 3nm + m^2 & \text{if } m \leq \sqrt{M/3}, \\ 2\frac{nm^2}{\sqrt{M/3}} + 2nm & \text{if } m \geq \sqrt{M/3}. \end{cases} \tag{2.2}$$

The bound for matrix multiplication holds for all values of $n \geq m$. The analysis here assumes the use of a conventional triangular solver and matrix multiplication, rather than so-called "fast" or Strassen-like algorithms. The asymptotic bounds for fast matrix-multiplication algorithms are better [5].

We analyze the recursively partitioned algorithm using induction. Initially, the analysis that does not take into account the permutation of rows that the algorithm performs. We shall return to these permutations later in this section. The recurrence that governs the total number of words that are transferred by the algorithm is

$$\begin{aligned} \mathbf{IO_{RP}}(n,1) &= 2n \ , \\ \mathbf{IO_{RP}}(n,m) &= \mathbf{IO_{RP}}(n,m/2) + \mathbf{IO_{RP}}(n-m/2,m/2) \\ &\quad + \mathbf{IO_{TS}}(m/2,m/2) \\ &\quad + \mathbf{IO_{MM}}(n-m/2,m/2,m/2) \ . \end{aligned}$$

We first prove by induction that if $1/2 \leq m/2 \leq \sqrt{M/3}$, then $\mathbf{IO_{RP}}(n,m) \leq 2nm(1+\lg m)$. The base case $m = 1$ is true. Assuming that the claim is true for $m/2$, for $m \geq 2$ we have

$$\begin{aligned} \mathbf{IO_{RP}}(n,m) &\leq 2n\frac{m}{2}\lg m + 2(n-m/2)\frac{m}{2}\lg m \\ &\quad + \frac{2.5m^2}{4} \\ &\quad + \frac{3nm}{2} - \frac{3m^2}{4} + \frac{m^2}{4} \\ &\leq 2nm\lg m + \frac{3nm}{2} + (0.5 - 2\lg m)\frac{m^2}{4} \\ &\leq 2nm(1+\lg m) \ . \end{aligned}$$

We now prove by induction that

$$\mathbf{IO_{RP}}(n,m) \leq 2nm\left(\frac{m}{2\sqrt{M/3}} + \lg m\right)$$

for $m/2 \geq \sqrt{M/3}$. The claim is true for the base case $m/2 = \sqrt{M/3}$ since $m/2 \leq \sqrt{M/3}$ and since $m/(2\sqrt{M/3}) = 1$. Assuming that the claim is true for $m/2$, we have

$$\begin{aligned}
\mathbf{IO_{RP}}(n,m) &\leq 2nm\left(\frac{m}{4\sqrt{M/3}} + \lg m - 1\right) \\
&\quad + \frac{m^3}{8\sqrt{M/3}} + \frac{m^2}{4} \\
&\quad + \frac{2(n-m/2)m^2}{4\sqrt{M/3}} + \frac{2(n-m/2)m}{2} \\
&\leq 2nm\left(\frac{m}{4\sqrt{M/3}} + \lg m - 1\right) \\
&\quad + \frac{m^3}{8\sqrt{M/3}} + \frac{m^2}{4} \\
&\quad + \frac{nm^2}{2\sqrt{M/3}} - \frac{m^3}{4\sqrt{M/3}} + nm - \frac{m^2}{2} \\
&\leq 2nm\left(\frac{m}{4\sqrt{M/3}} + \lg m - 1\right) \\
&\quad + \frac{nm^2}{2\sqrt{M/3}} - \frac{m^3}{8\sqrt{M/3}} + nm - \frac{m^2}{4} \\
&\leq 2nm\left(\frac{m}{4\sqrt{M/3}} + \lg m\right) \\
&\quad + \frac{nm^2}{2\sqrt{M/3}} \\
&= 2nm\left(\frac{m}{2\sqrt{M/3}} + \lg m\right) \ .
\end{aligned}$$

To bound the number word transfers due to permutations we compute the number of permutations a column undergoes during the algorithm. Each column is permuted either in the factorization in step 2 and in the permutation in step 7, or in the permutation in step 3 and in the factorization in step 6. It follows that each column is permuted $1 + \lg m$ times. If each word is brought from secondary memory, then the total number of I/Os required for permutations is at most $2n^2(1 + \lg m)$. This bound can be achieved when $n < M$ by reading entire columns to primary memory and permuting them in primary memory.

The following theorem summarizes the main result of this section.

THEOREM 2.1. *Given a matrix multiplication subroutine whose I/O performance satisfies equation* (2.2) *and a subroutine for solving triangular linear systems whose I/O performance satisfies equation* (2.1)*, the recursively partitioned LU decomposition algorithm running on a computer with M words of primary memory computes the LU decomposition with partial pivoting of an n-by-m matrix using at most*

$$\mathbf{IO_{RP}}(n,m) \leq 2nm\left(\frac{m}{2\sqrt{M/3}} + \lg m\right) + 2n^2(1 + \lg m)$$

I/Os. □

3. Analysis of the right-looking LU factorization. To put the performance of the recursively partitioned algorithm in perspective, we now analyze the performance of the column-block right-looking algorithm. We first describe the algorithm and then analyze the number of data transfers, or I/Os, it performs. While the bounds we obtain are asymptotically tight, we focus on lower bounds in terms of the constants. The number of I/Os required during the solution of triangular linear systems is smaller than the number of I/Os required during the updates to the trailing submatrix (a rank-r update to a matrix), so we ignore the triangular solves in the analysis.

Right-looking LU. The algorithm factors an n-by-m matrix A such that $PA = LU$, where $n \geq m$. The algorithm factors r columns in every iteration. In the kth iteration we decompose A into

$$PA = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix},$$

where A_{11} is a square matrix of order $(k-1)r$ and A_{22} is a square matrix of order r. In the kth iteration the algorithm performs the following steps.

1. Factor
$$P_2 \begin{bmatrix} A_{22} \\ A_{32} \end{bmatrix} = \begin{bmatrix} L_{22} \\ L_{32} \end{bmatrix} U_{22}.$$
2. Permute
$$\begin{bmatrix} A_{23} \\ A_{33} \end{bmatrix} \leftarrow P_2 \begin{bmatrix} A_{23} \\ A_{33} \end{bmatrix}.$$
3. Permute
$$\begin{bmatrix} L_{21} \\ L_{31} \end{bmatrix} \leftarrow P_2 \begin{bmatrix} L_{21} \\ L_{31} \end{bmatrix}.$$
4. Solve the triangular system $L_{22}U_{23} = A_{23}$.
5. Update $A_{33} \leftarrow A_{33} - L_{32}U_{23}$.

The number of I/Os required to factor an n-by-r matrix using the column-by-column algorithm is

$$\frac{nr^2}{4} \leq \mathbf{IO_{CF}}(n,r) \leq \frac{nr^2}{2},$$

when $M \leq nr/2$, but only

$$\mathbf{IO_{CF}}(n,r) = 2nr$$

when $M \geq nr$. To simplify the analysis, we ignore the range of M in which more than half the matrix fits within primary memory but less than the entire matrix. (Using one level of recursion leads to $\Theta(nr)$ I/Os in this range.) We use the facts that for $r \leq s$

$$\mathbf{IO_{TS}}(r,s) = \begin{cases} 2rs + \frac{r^2}{2} & \text{if } r < \sqrt{M/3}, \\ \frac{r^2 s}{\sqrt{M/3}} + rs & \text{if } r \geq \sqrt{M/3}, \end{cases} \tag{3.1}$$

and that for $r \leq s \leq t$

$$\mathbf{IO_{MM}}(t,r,s) = \begin{cases} 2ts + rs + rt & \text{if } r < \sqrt{M/3}, \\ 2\frac{trs}{\sqrt{M/3}} + 2ts & \text{if } r \geq \sqrt{M/3}. \end{cases} \tag{3.2}$$

The bound $2ts + rs + rt$ is an underestimate when $M < rs$. We ignore this small slack in the analysis.

The number of I/Os the algorithm performs depends on the relation of r to the dimensions of the matrix and to the size of memory. If r is so small that $M \geq nr$, then the updates to the trailing submatrix dominate the number of I/Os the algorithm performs. The $(m/r) - 1$ updates to the trailing submatrix require at least

$$\Theta(nm^2/r) = \Omega(n^2m^2/M)$$

I/Os. In particular, the first $m/2r$ updates require at least

$$\frac{m}{2r} 2\left(n - \frac{m}{2}\right)\frac{m}{2} \geq \frac{nm}{M}\left(n - \frac{m}{2}\right)\frac{m}{2} = \frac{n^2m^2}{2M} - \frac{nm^3}{4M} \geq \frac{n^2m^2}{4M} \ .$$

If r is larger, factoring the m/r blocks of r columns requires at least

$$\frac{m}{r}\frac{nr^2}{4} = \frac{nmr}{4}$$

I/Os. The number of I/Os required for the rank-r updates depends on the value of r. If $M/n \leq r \leq \sqrt{M/3}$, then the total number of I/Os performed by the rank-r updates is at least

$$\frac{m}{2r} 2\left(n - \frac{m}{2}\right)\frac{m}{2} \ .$$

Therefore, the number of I/Os performed by the algorithm is at least

$$\frac{nmr}{4} + \frac{m}{2r} 2\left(n - \frac{m}{2}\right)\frac{m}{2} \ ,$$

which is minimized at

$$r_{\text{opt}} = \sqrt{2m - m^2/n} \ .$$

For $n \geq m$, the optimal value of r lies between

$$\sqrt{m} \leq r_{\text{opt}} \leq \sqrt{2m} \ .$$

(The exact value might deviate slightly from this range, since the expression we derived for the number of I/Os is only a lower bound.) Substituting the optimal value of r, we find that the algorithm performs at least

$$\left(\frac{1}{4} + \frac{1}{2\sqrt{2}}\right) nm^{1.5} - \frac{1}{4\sqrt{2}} m^{2.5} \geq \left(\frac{1}{4} + \frac{1}{4\sqrt{2}}\right) nm^{1.5}$$

I/Os in this range. If $\sqrt{m} < M/n$, then the value $r = M/n$ yields better performance than $\sqrt{m}$. If $\sqrt{m} > \sqrt{M/3}$, then the value $r = \sqrt{M/3}$ yields better performance than $\sqrt{m}$.

If r is yet larger, $r \geq \sqrt{M/3}$, then the rank-r updates require

$$\Theta((m/r)nmr/\sqrt{M/3}) = \Theta(nm^2/\sqrt{M/3})$$

I/Os. In particular, the first $m/2r$ updates require at least

$$\frac{m}{2r}\frac{2(n-m/2)(m/2)r}{\sqrt{M/3}} \geq \frac{nm^2}{2\sqrt{M/3}} - \frac{m^3}{4\sqrt{M/3}} \geq \frac{nm^2}{4\sqrt{M/3}}$$

I/Os. The total number of I/Os in this range, including both the updates and the factoring of blocks of columns, is therefore at least

$$\frac{nm^2}{4\sqrt{M/3}} + \frac{nmr}{4}$$

if $r \geq \sqrt{M/3}, M/n$. The number of I/Os is minimized by choosing the smallest possible r, $r_{\text{opt}} = \sqrt{M/3}$.

If the matrix is not very large compared with the size of main memory, $n^2/3 \leq M$, it is also possible to choose r such that $\sqrt{M/3} \leq r \leq M/n$. In this case, the total number of I/Os is at least

$$\frac{nm^2}{4\sqrt{M/3}} + 2nm \geq \frac{n^2m^2}{4M} + 2nm \ .$$

The analysis can be summarized as follows. A value of r close to $\max(M/n, \sqrt{m})$ is optimal for almost all cases. The only exception is for truly huge matrices, where $M/3 \leq m$. For such matrices, $r = \sqrt{M/3}$ is better than $r = \sqrt{m}$. Combining the results, we obtain the following theorem.

THEOREM 3.1. *Given a matrix multiplication subroutine whose I/O performance satisfies equation* (3.1) *and a subroutine for solving triangular linear systems whose I/O performance satisfies equation* (3.2)*, the right-looking LU decomposition algorithm running on a computer with M words of primary memory computes the LU decomposition with partial pivoting of an n-by-m matrix using at least*

$$\mathbf{IO_{RL}}(n,m) \geq \begin{cases} \frac{1}{4}\frac{n^2m^2}{M} & \text{if } r = M/n, \\ \frac{1}{4}nm^{1.5} & \text{if } r \approx \sqrt{m} < \sqrt{M/3}, \\ \frac{1}{4}nm^{1.5} & \text{if } r = \sqrt{M/3} \end{cases}$$

I/Os. □

The first case, $r = M/n$, leads to better performance only when more than $\sqrt{m}$ columns fit within primary memory. Although these are lower bounds, they are asymptotically tight. The value 1/4 is a lower bound on the actual constant, which is higher than that.

4. Experimental results. We have implemented and tested the recursively partitioned algorithm[1]. The goal of the experiments was to determine whether the recursively partitioned algorithm is more efficient than the right-looking algorithm in

[1] Our Fortran 90 implementation is available online by anonymous ftp from theory.lcs.mit.edu as /pub/people/sivan/dgetrf90.f. The code can be compiled by many Fortran 77 compilers, including compilers from IBM, Silicon Graphics, and Digital, by removing the RECURSIVE key word and using a compiler option that enables recursion (see [11] for details).

practice. The results of the experiments clearly show that the recursively partitioned algorithm performs less I/O and is that it is faster, at least on the computer on which the experiments were conducted.

The results of the experiments complement our analysis of the two algorithms. The analysis shows that the recursively partitioned algorithm performs less I/O than the right-looking algorithm for most values of n and M. The analysis stops short of demonstrating that one algorithm is faster than another in three respects. First, the bounds in the analysis are not exact. Second, the analysis counts the total number of I/Os in the algorithm, but the distribution of the I/O within the algorithm is significant. Finally, the analysis uses a simplified model of a two-level hierarchical memory that does not capture all the subtleties of actual memory systems. The experiments show that even though our analysis is not exact in these respects, the recursively partitioned algorithm is indeed faster.

Three sets of experiments are presented in this section. The first set presents and analyzes in detail experiments on IBM RS/6000 workstations. The goal of this set of experiments is to establish that the recursively partitioned algorithm is faster than the right-looking algorithm. The second set of experiments show, in less detail, that the recursively partitioned algorithm outperforms LAPACK's right-looking algorithm on a wide range of architectures. The goal of the second set of experiments is to establish the robustness of the performance of the recursively partitioned algorithm. The third set of experiments shows that using Strassen's matrix-multiplication algorithm speeds up the recursively partitioned algorithm but does not seem to speed up the right-looking algorithm.

Some of the technical details of the experiments, such as operating system versions, compiler versions, and compiler options are omitted from this paper. These details are fully described in our technical report [11].

Detailed experimental analyses. The first set of experiments was performed on an IBM RS/6000 workstation with a 66.5 MHz POWER2 processor [14], 128 Kbytes 4-way set associative level-1 data cache, a 1 Mbyte direct-mapped level-2 cache, and a 128-bit-wide main memory bus. The POWER2 processor is capable of issuing two double-precision multiply-add instructions per clock cycle. Both LAPACK's right-looking LU-factorization subroutine DGETRF and the recursively partitioned algorithm were compiled by IBM's XLF compiler version 3.2. All the algorithms used the BLAS from IBM's Engineering and Scientific Subroutine Library (ESSL). On square matrices we have also measured the performance of the LU-factorization subroutine DGEF from ESSL. The interface of this subroutine only allows for the factorization of square matrices. The coding style and the data structures used in the recursively partitioned algorithm are the same as the ones used by LAPACK. In particular, permutations are represented in both algorithms as a sequence of exchanges. In all cases, the array that contains the matrix to be factored was allocated statically and aligned on a 16-byte boundary. The leading dimension of the matrix was equal to the number of rows (no padding).

The performance of the algorithms was assessed using measurements of both running time and cache misses. Time was measured using the machines real-time clock, which has a resolution of one cycle. The number of cache misses was measured using the POWER2 performance monitor [13]. The performance monitor is a hardware subsystem in the processor capable of counting cache misses and other processor events. Both the real-time clock and the performance monitor are oblivious to time sharing. To minimize the risk that measurements are influenced by other processes, we ran

TABLE 4.1

The performance in millions of operations per second (Mflops) and the number of cache misses per thousand floating point operations (CM/Kflop) of five LU-factorization algorithms on an IBM RS/6000 workstation, on square matrices. The figures for LAPACK's DGETRF are those of the block size r with the best running time, in upright letters, and those of the block size with the smallest number of cache misses, in italics. The minimum number of cache misses does not generally coincide with the minimum running time. See the text for a full description of the experiments.

	$n = 1007$		$n = 1024$	
Subroutine	Mflops	CM/Kflop	Mflops	CM/Kflop
LAPACK's DGETRF, row exchanges	178, *176*	5.81, *5.65*	170, *168*	5.45, *5.29*
Recursively partitioned, row exchanges	201	3.76	186	4.14
LAPACK's DGETRF, permuting by columns	201, *199*	2.94, *2.81*	198, *195*	3.11, *3.02*
Recursively partitioned, permuting by columns	222	1.61	223	1.59
ESSL's DGEF	228	2.15	221	3.42

the experiments when no other users used the machine (but it was connected to the network). We later verified that the measurements are valid by comparing the real-time-clock measurements with the user time reported by AIX's getrusage system call on an experiment-by-experiment basis. All measurements reported here are based on an average of 10 executions.

We have coded two variants of the recursively partitioned algorithm. The two versions differ in the way permutations are applied to submatrices. In one version, permutations are applied using LAPACK's auxiliary subroutine DLASWP. This subroutine, which is also used by LAPACK's right-looking algorithm, permutes the rows of a submatrix by exchanging rows using the vector exchange subroutine DSWAP, a level-1 BLAS. The second version permutes the rows of the matrix by applying the entire sequence of exchanges to one column after another. The difference amounts to swapping the inner and outer loops. This change was suggested by Fred Gustavson.

The first experiment, whose results are summarized in Table 4.1, was designed to determine the effects of a complex hierarchical memory system on the partitioned algorithms. Four facts emerge from the table.

1. The recursively partitioned algorithm performs fewer cache misses and delivers higher performance than the right-looking algorithm. ESSL's subroutine performs less cache misses than LAPACK but more than the recursively partitioned algorithm, but it achieves best or close to best performance.
2. Permuting one column at a time leads to fewer cache misses and faster execution than exchanging rows. This is true for both the right-looking algorithm and the recursively partitioned algorithm. This is probably a result of the advantage of the stride-1 access to the column in the column permuting over the large stride access to rows in the row exchanges.
3. The performance, in terms of both time and cache misses, of all the algorithms except the recursively partitioned with column permuting is worse when the leading dimension of the matrix is a power of 2 than when it is not. The performance of the recursively partitioned algorithm with column permuting improves by less than half a percent. The degradation in performance on a power of 2 is probably caused by fact that the caches are not fully associative.
4. The running time depends on the measured number of cache misses but not completely. This can be seen both from the fact that ESSL's DGEF performs more cache misses than the recursively partitioned algorithm, but it is faster, and from the fact that the block size that leads to the minimum

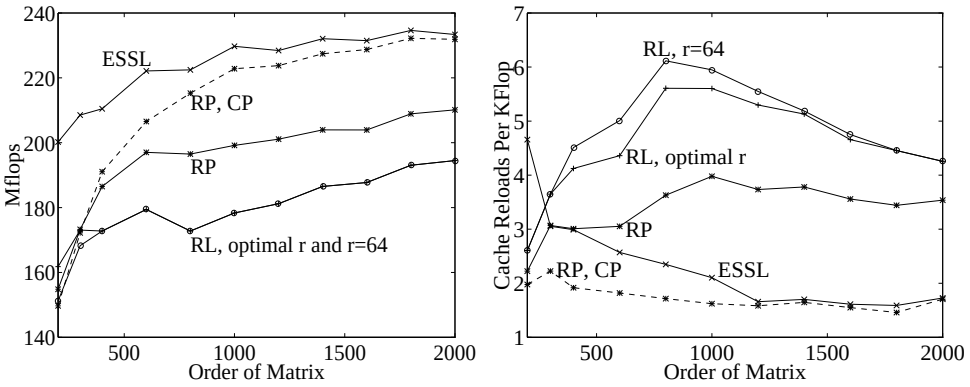


FIG. 4.1. *The performance in Mflops* (*on the left*) *and the number of cache misses per Kflop* (*on the right*) *of LU factorization algorithms on an IBM RS/*6000 *workstation. These graphs depict the performance of the recursively partitioned (RP) and right-looking (RL) algorithms on square matrices. The optimal value of* r *was selected experimentally from powers of* 2 *between* 2 *and* 256. *The dashed lines represent the performance of the recursively partitioned algorithms with column permuting* (*CP*).

number of cache misses in the DGETRF does not lead to the best running time. The discrepancy can be caused by several factors that are not measured, including misses and conflicts in the level-2 cache, TLB misses, and instruction scheduling. In all four cases in the table the minimum running time is achieved with a value of r that is higher than the number that leads to a minimum number of cache misses. For example, on $n = 1007$, DGETRF with row exchanges performed the least number of cache misses with $r = 40$, but the fastest running time was achieved with $r = 55$. This may mean that the cause of the discrepancy is misses in the level-2 cache, which is larger than the level-1 cache and therefore may favor a larger block size (since more columns fit in it).

In summary, the experiment shows that although the implementation details of the memory system influence the performance of the algorithms, the recursively partitioned algorithm still emerges as faster than the right-looking one when they are implemented in a similar way.

The second set of experiments was designed to assess the performance of the algorithms over a wide range of input sizes. The performance and number of cache misses of the algorithms are presented in Figure 4.1 on square matrices ranging in order from 200 to 2000. The level-1 cache is large enough to store a matrix of order 128. The following points emerge from the experiment.

1. Beginning with matrices of order $n = 300$, the recursively partitioned algorithm with column permuting is faster than the same algorithm with row exchanges, which is still faster than LAPACK's DGETRF with row exchanges (we did not measure the performance of DGETRF with column permuting in this experiment).
2. The performance of DGETRF with optimal block size r and with $r = 64$ is essentially the same except at $n = 300$, although the optimal block size clearly leads to a smaller number of cache misses from $n = 400$ through $n = 1600$.
3. The recursively partitioned algorithm performs fewer cache misses than ESSL's DGEF on all input sizes, but it is not faster. As in the first experiment, the

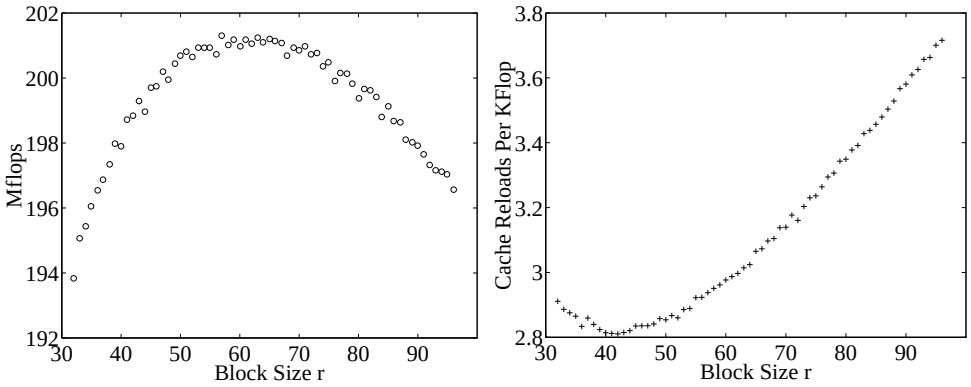


FIG. 4.2. *The performance in Mflops* (*on the left*) *and the number of cache misses per Kflop* (*on the right*) *of the right-looking algorithm with column permuting with as a function of the block size* r. *The order of the square matrix used is* $n = 1007$. *Note that the y-axes do not start from zero.*

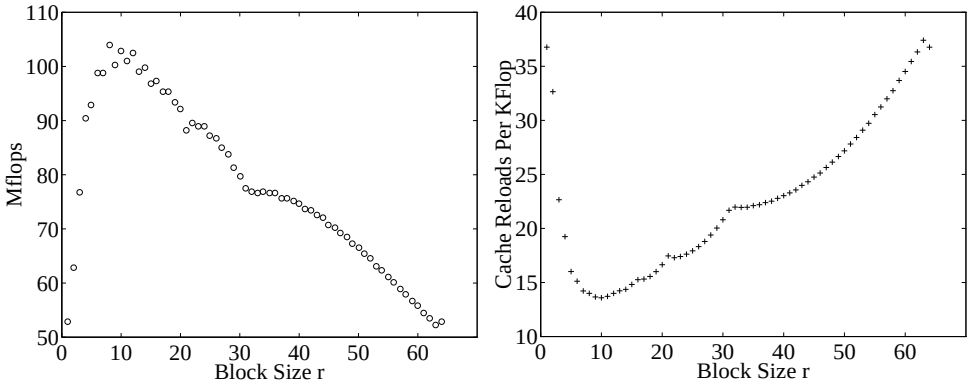


FIG. 4.3. *The performance in Mflops* (*on the left*) *and the number of cache misses per Kflop* (*on the right*) *of the right-looking algorithm with column permuting with as a function of the block size* r. *The dimensions of the matrix are* 62500-*by*-64. *For comparison, the performance of the recursively partitioned algorithm on this problem is* 118 *Mflops and* 11.03 *CM/Kflop.*

experiment itself does not indicate what causes this phenomenon. We speculate that it is caused by better instruction scheduling or fewer misses in the level-2 cache.

The next experiment was designed to determine the sensitivity of the performance of the right-looking algorithm to the block size r. We used the column permuting strategy which proved more efficient in the previous experiments. The experiment consists of running the algorithm on a range of block sizes on a square matrix of order 1007 and on a rectangular 62500-by-64 matrix. The factorization of a rectangular matrix with $n > m$ arises as a subproblem in out-of-core LU factorization algorithms that factor blocks of columns that fit within core. The specific dimensions of the matrices were chosen so as to minimize the effects of conflicts in the memory system on the results. The results for $n = m = 1007$, shown in Figure 4.2, show that the minimum number of cache misses occurs at $r = 42$, which is higher than $\sqrt{m} \approx 32$, and that the best performance is achieved with an even higher value of r, 55. The

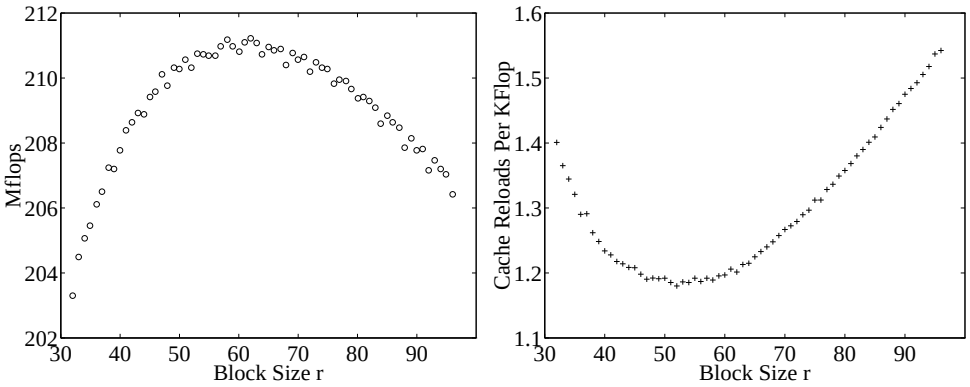


FIG. 4.4. *The performance in Mflops* (*on the left*) *and the number of cache misses per Kflop* (*on the right*) *of the right-looking algorithm with column permuting with as a function of the block size* r. *The order of the square matrix used is* $n = 1007$. *The machine used here has a bigger level-*1 *cache and no level-*2 *cache than the machine used in all the other experiments. Compare with Figure* 4.2. *For comparison, the performance of the recursively partitioned algorithm on this problem on this machine is* 229 *Mflops and* 0.650 *CM/Kflop.*

performance is not very sensitive to the choice of r, however, and all values between about 50 and 70 yield essentially the same performance, 201 Mflops. The results for 62500-by-64 matrices, shown in Figure 4.3, show that the minimum number of cache misses occur at $r = 10$, and the best performance occurs at $r = 8$, which happens to coincide exactly with $\sqrt{m}$. The sensitivity to r is greater here than in the square case, especially below the optimal value.

The last experiment in this set, presented in Figure 4.4, was designed to determine whether the discrepancy between the optimal block size in terms of level-1 cache misses and the optimal block size in terms of running time was caused by the level-2 cache. The experiment repeats the last experiment for square matrices of order 1007, except that the experiment was conducted on a machine with a 256-bit-wide main memory bus, 256 Kbytes level-1 cache, and no level-2 cache. The two machines are identical in all other respects. There is a discrepancy in optimal block sizes in Figure 4.4, but it is smaller than the discrepancy in Figure 4.2. The experiment shows that the discrepancy is not caused solely by the level-2 cache. It is not possible to determine whether the smaller discrepancy in this experiment is due to the lack of level-2 cache or to the larger level-1 cache.

Robustness experiments. The second set of experiments shows that the performance advantage of the recursively partitioned algorithm, which was demonstrated by the first set of experiments, is not limited to a single computer architecture. The experiments accomplish this goal by showing that the recursively partitioned algorithm outperforms the right-looking algorithm on a wide range of architectures.

All the experiments in this set compare the performance of the recursively partitioned algorithm with the performance of LAPACK's right-looking algorithm on two sizes of square matrices, $n = 1007$ and $n = 2014$ (except when the larger matrices do not fit within main memory). These sizes were chosen so as to minimize the impact of cache associativity on the results. Each measurement reported represents the average of the best five out of 10 runs, to minimize the effect of other processes in the system. The block size for the right-looking algorithm was LAPACK's default $r = 64$.

TABLE 4.2

The running time in seconds of LU factorization algorithms on several machines. For each machine and each matrix order, the table shows the running times of the recursively partitioned (RP) algorithm and the right-looking (RL) algorithm with row exchanges and column permutations. Some measurements are not available and marked as N/A because the amount of main memory is insufficient to factor the larger matrix in core. See the text for a full description of the experiments.

	$n = 1007$				$n = 2014$			
	Row exchanges		Column pivoting		Row exchanges		Column pivoting	
Machine	RL	RP	RL	RP	RL	RP	RL	RP
IBM POWER2	3.82	3.39	3.37	3.05	27.88	26.00	25.28	23.45
IBM POWER	22.81	19.07	17.87	16.86	146.1	143.4	135.8	129.7
SGI R4600/R4610	37.15	34.39	36.42	33.57	N/A	N/A	N/A	N/A
SGI R4400/R4010	9.44	8.36	9.38	8.29	73.92	68.64	73.73	66.79
DEC A21064	9.27	8.94	9.36	8.89	N/A	N/A	N/A	N/A
DEC A21164	2.61	2.56	2.30	2.25	20.15	19.68	17.80	17.06

We used the following machine configurations:

- A 66.5 MHz IBM RS/6000 workstation with a POWER2 processor, 128 Kbytes 4-way set associative data cache, a 1 Mbyte direct-mapped level-2 cache, and a 128-bit-wide bus. We used the BLAS from IBM's ESSL.
- A 25 MHz IBM RS/6000 workstation with a POWER processor, 64 Kbytes 4-way set associative data cache, and a 128-bit-wide bus. We used the BLAS from IBM's ESSL.
- A 100 MHz Silicon Graphics Indy workstation with a MIPS R4600/R4610 CPU/FPU pair, a 16 Kbytes direct-mapped data cache, and a 64-bit-wide bus. We used the SGI BLAS. This machine has only 32 Mbytes of main memory, so the experiment does not include matrices of order $n = 2014$.
- A 250 MHz Silicon Graphics Onyx workstation with 4 MIPS R4400/R4010 CPU/FPU pairs, a 16 Kbytes direct-mapped data cache per processor, a 4 Mbytes level-2 cache per processor, and a 2-way interleaved main memory system with a 256-bit-wide bus. The experiment used only one processor. We used the SGI BLAS.
- A 150 MHz DEC 3000 Model 500 with an Alpha 21064 processor, 8 Kbytes direct-mapped cache, and a 512 Kbytes level-2 cache. We used the BLAS from DEC's DXML for IEEE floating point. A limit on the amount of physical memory allocated to a process prevented us from running the experiment on matrices of order $n = 2014$.
- A 300 MHz Digital AlphaServer with 4 Alpha 21164 processors, each with an 8 Kbytes level-1 data cache, a 96 Kbytes on-chip level-2 cache, and a 4 Mbytes level-2 cache. The experiment used only one processor. We used the BLAS from DEC's DXML for IEEE floating point.

The results, which are reported in Table 4.2, show that the recursively partitioned algorithm consistently outperforms the right-looking algorithm. The results also show that permuting columns is almost always faster than exchanging rows.

Experiments using Strassen's algorithm. Performing the updates of the trailing submatrix using a variant of Strassen's algorithm [10] improved the performance of the recursively partitioned algorithm. We replaced the call to DGEMM, the level-3 BLA subroutine for matrix multiply-add by a call to DGEMMB, a pub-

lic domain implementation[2] of a variant of Strassen algorithm [3]. (Replacing the calls to DGEMM with calls to a Strassen matrix-multiplication subroutine in IBM's ESSL gave similar results.) DGEMMB uses Strassen's algorithm only when all the dimensions of the input matrices are greater than a machine-dependent constant. The authors of DGEMMB set this constant to 192 for IBM RS/6000 workstations.

In the recursively partitioned algorithm with column permuting, the replacement of DGEMM by DGEMMB reduced the factorization time on the POWER2 machine to 2.99 seconds for $n = 1007$ and to 22.18 seconds for $n = 2014$. The factorization times with the conventional matrix-multiplication algorithm, reported in the first line of Table 4.2, are 3.05 and 23.45 seconds. The running time was reduced from 182.7 to 166.8 seconds on a matrix of order $n = 4028$. The change would have no effect on the right-looking algorithm, since in all the matrices it multiplies at least one dimension is r which was smaller than 192 in all the experiments.

A similar experiment carried out by Bailey, Lee, and Simon [2] showed that Strassen's algorithm can accelerate the LAPACK's right-looking LU factorization on a Cray Y-MP. The largest improvements in performance, however, occurred when large values of r were used. The fastest factorization of a matrix of order $n = 2048$, for example, was obtained with $r = 512$. Such a value is likely to cause poor performance on machines with caches. (The Cray Y-MP has no cache.) On the IBM POWER2 machine, which has caches, increasing r from 64 to 512 causes the factorization time with a conventional matrix-multiplication algorithm to increase from 30.8 seconds to 54 seconds. Replacing the matrix-multiplication subroutine by DGEMMB with $r = 512$ reduces the solution time but by less than two seconds.

5. Conclusions. The recursively partitioned algorithm should be used instead of the right-looking algorithm because it delivers similar or better performance without parameters that must be tuned. No parameter to choose means that there is no possibility of a poor choice, and hence the new algorithm is more robust. Section 4 shows that the performance of the right-looking algorithm can be sensitive to r and that the best performance does not always coincide with the block size that causes the smallest number of cache misses. Choosing r can be especially difficult on machines with more than two levels of memory. A recursive algorithm, on the other hand, is a natural choice for hierarchical memory systems with more than two levels.

The recursively partitioned algorithm provides a good opportunity to use a fast matrix-multiplication algorithm such as Strassen's algorithm. Since a significant fraction of the work performed by the recursively partitioned algorithm is used to multiply large matrices, the benefit of using Strassen's algorithm can be large. The right-looking algorithm performs the same work by several multiplications of smaller matrices, so the benefit of Strassen's algorithm should be smaller.

The analysis of the right-looking algorithm in section 3 shows how the block size r should be chosen. The value $r \approx \sqrt{m}$ is optimal with two exceptions. When a single row is too large to fit within primary memory, a value $r = \sqrt{M/3}$ leads to better performance. When more than $\sqrt{m}$ columns fit within primary memory, r should be set to M/n to minimize memory traffic. The extreme cases are the source of the difficulty in choosing a good value of r for hierarchical memory systems with more than two levels. In our experiments, the performance of the right-looking algorithm on matrices with more rows than columns was very sensitive to the choice of r, but it was not sensitive on large square matrices.

[2]Available online from http://www.netlib.org/linalg/gemmw.

In the typical cases, when at least one row fits within primary memory, the right-looking algorithm with an optimal choice of r performs a factor of $\Theta(\sqrt{M/m})$ more data transfers than the recursively partitioned algorithm. In our experiments this factor led to a significant difference in both the number of cache misses and the running time.

The conclusion that the value $r = \sqrt{m}$ is often close to optimal shows that there is a system-independent way to choose r. In comparison, the model implementation of ILAENV, LAPACK's block-size-selection subroutine, uses a fixed value, 64, and LAPACK's *User's Guide* advises that system-dependent tuning of r could improve performance. The viewpoint of the LAPACK designers seems to be that r is a system-dependent parameter whose role is to hide the low bandwidth of the secondary memory system during the updates of the trailing submatrices. Our analysis here shows that the true role of r is to balance the number of data transfers between the two components of the algorithm: the factorization of blocks of columns and the updates of the trailing submatrices.

Designers of out-of-core LU decomposition codes often propose to use block-column (or row) algorithms. Many of them propose to choose $r = M/n$ so that an entire block of columns fits within primary memory [4, 6, 7, 15]. This approach works well when the columns are short and a large number of them fit within primary memory, but the performance of such algorithms would be unacceptable when only few columns fit within primary memory. Some researchers [7, 8, 9] suggest that algorithms that use less primary memory than is necessary for storing a few columns might have difficulty implementing partial pivoting. The analysis in this paper shows that it is possible to achieve a low number of data transfers even when a single row or column does not fit within primary memory.

Womble et al. [15] presented a recursively partitioned LU decomposition algorithm without pivoting. They claimed, without a proof, that pivoting can be incorporated into the algorithm without asymptotically increasing the number of I/Os the algorithm performs. They suggested that a recursive algorithm would be difficult to implement, so they implemented instead a partitioned left-looking algorithm using $r = M/n$.

Toledo and Gustavson [12] describe a recursively partitioned algorithm for out-of-core LU decomposition with partial pivoting. Their algorithm uses recursion on large submatrices but switches to a left-looking variant on smaller submatrices (that would still not fit within main memory). Depending on the size of main memory, their algorithm can factor a matrix in 2/3 the amount of time used by an out-of-core left-looking algorithm with a fixed block size.

Acknowledgments. Thanks to Rob Schreiber for reading several early versions of this paper and commenting on them. Thanks to Fred Gustavson and Ramesh Agarwal for helpful suggestions. Thanks to the anonymous referees for several helpful comments.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. D. Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK User's Guide*, SIAM, Philadelphia, PA, 2nd ed., 1994. Also available online from http://www.netlib.org.

[2] D. H. Bailey, K. Lee, and H. D. Simon, *Using Strassen's algorithm to accelerate the solution of linear systems*, J. of Supercomputing, 4 (1990), pp. 357–371.

[3] C. C. Douglas, M. Heroux, G. Slishman, and R. M. Smith, *GEMMW: A portable level* 3 *BLAS Winograd variant of Strassen's matrix-matrix multiply algorithm*, J. Comput. Phys., 110 (1994), pp. 1–10.

[4] J. J. Du Cruz, S. M. Nugent, J. K. Reid, and D. B. Taylor, *Solving large full sets of linear equations in a paged virtual store*, ACM Trans. Math. Software, 7 (1981), pp. 527–536.

[5] P. C. Fischer and R. L. Probert, *A note on matrix multiplication in a paging environment*, in ACM '76: Proceedings of the Annual Conference, Houston, TX, 1976, pp. 17–21.

[6] N. Geers and R. Klees, *Out-of-core solver for large dense nonsymmetric linear systems*, Manuscripta Geodetica, 18 (1993), pp. 331–342.

[7] R. G. Grimes, *Solving systems of large dense linear equations*, J. of Supercomputing, 1 (1988), pp. 291–299.

[8] A. C. McKeller and E. G. Coffman, Jr., *Organizing matrices and matrix operations for paged memory systems*, Communications of the ACM, 12 (1969), pp. 153–165.

[9] C. B. Moler, *Matrix computations with Fortran and paging*, Communications of the ACM, 15 (1972), pp. 268–270.

[10] V. Strassen, *Gaussian elimination is not optimal*, Numer. Math., 13 (1969), pp. 354–355.

[11] S. Toledo, *Locality of Reference in LU Decomposition with Partial Pivoting*, Tech. report RC20344, IBM T.J. Watson Research Center, Yorktown Heights, NY, Jan. 1996.

[12] S. Toledo and F. G. Gustavson, *The design and implementation of SOLAR, a portable library for scalable out-of-core linear algebra computations*, in Proceedings of the 4th Annual Workshop on I/O in Parallel and Distributed Systems, Philadelphia, PA, May 1996, pp. 28–40.

[13] E. H. Welbon, C. C. Chan-Nui, D. J. Shippy, and D. A. Hicks, *The POWER*2 *performance monitor*, IBM J. Res. Develop., 38 (1994), pp. 545–554.

[14] S. W. White and S. Dhawan, *POWER*2*: Next generation of the RISC System/*6000 *family*, IBM J. Res. Develop., 38 (1994).

[15] D. Womble, D. Greenberg, S. Wheat, and R. Riesen, *Beyond core: Making parallel computer I/O practical*, in Proceedings of the 1993 DAGS/PC Symposium, Hanover, NH, June 1993, Dartmouth Institute for Advanced Graduate Studies, pp. 56–63. Also available online from http://www.cs.sandia.gov/~dewombl/parallel_io_dags93.html.

SIAM J. MATRIX ANAL. APPL.
Vol. 18, No. 4, pp. 1082–1092, October 1997

017

ON A VARIATIONAL FORMULATION OF THE GENERALIZED SINGULAR VALUE DECOMPOSITION*

MOODY T. CHU†, ROBERT. E. FUNDERLIC‡, AND GENE H. GOLUB§

Abstract. A variational formulation for the generalized singular value decomposition (GSVD) of a pair of matrices $A \in R^{m\times n}$ and $B \in R^{p\times n}$ is presented. In particular, a duality theory analogous to that of the SVD provides new understanding of left and right generalized singular vectors. It is shown that the intersection of row spaces of A and B plays a key role in the GSVD duality theory. The main result that characterizes left GSVD vectors involves a generalized singular value deflation process.

Key words. generalized eigenvalue and eigenvector, generalized singular value and singular vector, stationary value and stationary point, deflation, duality

AMS subject classifications. 65F15, 65H15

PII. S0895479895287079

1. Introduction.

The singular value decomposition (SVD) of a given matrix $A \in R^{m\times n}$ is

$$U^T AV = S = \text{diag}\{\sigma_1, \ldots, \sigma_q\},\ q = \min\{m, n\}, \tag{1}$$

where $U \in R^{m\times m}$ and $V \in R^{n\times n}$ are orthogonal matrices, $S \in R^{m\times n}$ is zero except for the real nonnegative elements $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_q = 0$ on the leading diagonal with $r = \text{rank}(A)$. The $\sigma_i, i = 1, \ldots, q$, are the singular values of A. Of the many ways to characterize the singular values of A, the following variational property is of particular interest [4].

THEOREM 1.1. *Consider the optimization problem*

$$\max_{x\neq 0} \frac{\|Ax\|}{\|x\|}, \tag{2}$$

where $\|\cdot\|$ denotes the 2-norm of a vector. Then the singular values of A are precisely the stationary values, i.e., the functional evaluations at the stationary points, of the objective function $\|Ax\|/\|x\|$ with respect to $x \neq 0$.

We note that the stationary points $x \in R^n$ in problem (2) are the right singular vectors of A. At each of such points, it follows from the usual duality theory that there exists a vector $y \in R^m$ of unit Euclidean length such that $y^T Ax$ is equal to the corresponding stationary value. This y is the corresponding left singular vector of A.

The main purpose of this paper is to delineate a similar variational principle that leads to the generalized singular value decomposition (GSVD) of a pair of matrices

* Received by the editors May 30, 1995; accepted for publication (in revised form) by P. Van Dooren November 18, 1996.

http://www.siam.org/journals/simax/18-4/28707.html

† Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (chu@math.ncsu.edu). This research was supported in part by National Science Foundation grants DMS-9123448 and DMS-9422280.

‡ Department of Computer Science, North Carolina State University, Raleigh, NC 27695-8206 (ref@adm.csc.ncsu.edu).

§ Department of Computer Science, Stanford University, Stanford, CA 94305 (golub@sccm.stanford.edu). This research was supported in part by National Science Foundation grant CCR-9505393.

$A \in R^{m\times n}$ and $B \in R^{p\times n}$. While the variational formula analogous to (2) for the GSVD is well known, the corresponding duality theory has apparently not been developed. (See [1] for a related treatise.) The purpose of this note is to fill the duality theory gap for the GSVD problem.

Let $\mathcal{R}(M)$ and $\mathcal{N}(M)$ denote, respectively, the range space and the null space of any given matrix M. We will see that the intersection of row spaces of A and B,

$$\mathcal{R}(A^T)\bigcap\mathcal{R}(B^T) = \left\{z \in R^n \mid z^T = x^TA = y^TB \text{ for some } x \in R^m \text{ and } y \in R^p\right\},$$

plays a fundamental role in the duality theory of the associated GSVD. The equivalence

$$C\begin{bmatrix} x \\ -y \end{bmatrix} = [A^T, B^T]\begin{bmatrix} x \\ -y \end{bmatrix} = 0 \Longleftrightarrow x^TA = y^TB$$

suggests that the null space of the matrix $C := [A^T, B^T]$,

$$\mathcal{N}(C) = \left\{\begin{bmatrix} x \\ -y \end{bmatrix} \in R^{m+p} \mid C\begin{bmatrix} x \\ -y \end{bmatrix} = 0\right\}, \tag{3}$$

may be interpreted as a "representation" of $\mathcal{R}(A^T)\bigcap\mathcal{R}(B^T)$. But this representation is not unique in that different values of $\begin{bmatrix} x \\ -y \end{bmatrix} \in \mathcal{N}(C)$ may give rise to the same $z \in \mathcal{R}(A^T)\bigcap\mathcal{R}(B^T)$. In particular, all points in the subspace

$$\mathcal{S} := \left\{\begin{bmatrix} g \\ -h \end{bmatrix} \in R^{m+p} \mid g^TA = h^TB = 0\right\} \tag{4}$$

collapse into the zero vector in $\mathcal{R}(A^T)\bigcap\mathcal{R}(B^T)$. For a reason to be discussed in what follows (see (16) and the argument thereafter), the subspace $\mathcal{S}$ should be taken out of consideration. More precisely, define the *homomorphism* $H : \mathcal{N}(C) \longrightarrow \mathcal{R}(A^T)\bigcap\mathcal{R}(B^T)$ by

$$z = H\left(\begin{bmatrix} x \\ -y \end{bmatrix}\right) \Longleftrightarrow z^T = x^TA = y^TB,$$

and define, for every $\begin{bmatrix} x \\ -y \end{bmatrix} \in \mathcal{N}(C)$, the *quotient map* $\pi\left(\begin{bmatrix} x \\ -y \end{bmatrix}\right)$ to be the coset of $\mathcal{S}$ containing $\begin{bmatrix} x \\ -y \end{bmatrix}$, i.e.,

$$\pi\left(\begin{bmatrix} x \\ -y \end{bmatrix}\right) := \begin{bmatrix} x \\ -y \end{bmatrix} + S. \tag{5}$$

Then the first homomorphism theorem for vector spaces (see, for example, [5, Theorem 4.a]) states that $\mathcal{R}(A^T)\bigcap\mathcal{R}(B^T)$ is *isomorphic* to the quotient space $\mathcal{N}(C)/\mathcal{S}$ where

$$\mathcal{N}(C)/\mathcal{S} := \left\{\pi\left(\begin{bmatrix} x \\ -y \end{bmatrix}\right) \mid \begin{bmatrix} x \\ -y \end{bmatrix} \in \mathcal{N}(C)\right\}. \tag{6}$$

It is in this quotient space that we establish the duality theory.

Recall that linearly independent vectors in $\mathcal{N}(C)$ that are not in $\mathcal{S}$ will generate naturally linearly independent vectors in the quotient space $\mathcal{N}(C)/\mathcal{S}$ through the quotient map. Thus the simplest way to represent $\mathcal{N}(C)/\mathcal{S}$ is through the orthogonal complement $\mathcal{S}^{\perp}$ of $\mathcal{S}$ in $\mathcal{N}(C)$. Define $N(A^T)$ and $N(B^T)$ to be matrices so that their columns span, respectively, the null spaces $\mathcal{N}(A^T)$ and $\mathcal{N}(B^T)$. Define

$$Z := \begin{bmatrix} A^T & B^T \\ N(A^T)^T & 0 \\ 0 & N(B^T)^T \end{bmatrix}. \tag{7}$$

Then $\mathcal{N}(C)/\mathcal{S}$ can be uniquely represented by the subspace

$$\mathcal{N}(Z) = \left\{ \begin{bmatrix} x \\ -y \end{bmatrix} \in R^{m+p} \mid Z \begin{bmatrix} x \\ -y \end{bmatrix} = 0 \right\}. \tag{8}$$

We shall have the dimension counted carefully in section 2.

Our discussion is based upon the following formulation of the GSVD for A and B by Paige and Saunders [6] (or QSVD in [3]) that generalizes the original concept in [9].

DEFINITION 1.1. *Assume* $\operatorname{rank}(C) = k$, *then there exist orthogonal* $U \in R^{m\times m}$, $V \in R^{p\times p}$, *and invertible* $X \in R^{n\times n}$ *such that*

$$\begin{bmatrix} U^T & 0 \\ 0 & V^T \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} X = \begin{bmatrix} \Omega_A & 0 \\ \Omega_B & 0 \end{bmatrix} \tag{9}$$

with $\Omega_A \in R^{m\times k}$ *and* $\Omega_B \in R^{p\times k}$ *given by*

$$\Omega_A = \begin{array}{c} \begin{matrix} r & s & k-r-s \end{matrix} \\ \begin{bmatrix} I_A & & \\ & S_A & \\ & & O_A \end{bmatrix} \end{array} \begin{array}{l} r \\ s \\ m-r-s, \end{array}$$

$$\Omega_B = \begin{array}{c} \begin{matrix} r & s & k-r-s \end{matrix} \\ \begin{bmatrix} O_B & & \\ & S_B & \\ & & I_B \end{bmatrix} \end{array} \begin{array}{l} p-k+r \\ s \\ k-r-s, \end{array}$$

where I_A *and* I_B *are identity matrices,* O_A *and* O_B *are zero matrices with possibly no rows or columns, and* $S_A = \operatorname{diag}\{\omega_A^{(1)}, \ldots, \omega_A^{(s)}\}$ *and* $S_B = \operatorname{diag}\{\omega_B^{(1)}, \ldots, \omega_B^{(s)}\}$ *satisfy*

$$1 > \omega_A^{(1)} \geq \cdots \geq \omega_A^{(s)} > 0, \qquad 0 < \omega_B^{(1)} \leq \cdots \leq \omega_B^{(s)} < 1,$$
$$\omega_A^{(i)^2} + \omega_B^{(i)^2} = 1.$$

The quotients

$$\lambda_i := \frac{\omega_A^{(i)}}{\omega_B^{(i)}}, \qquad i = 1, \ldots, s,$$

are called the generalized singular values of (A, B) *for which we make use of the notation* $\Lambda := \mathrm{diag}\{\lambda_1, \ldots, \lambda_s\}$. *The values of* r *and* s *are defined internally by the matrices* A *and* B.

Suppose we partition X into four blocks of columns $X = [X_1, X_2, X_3, X_4]$ with column sizes $r, s, k-r-s$, and $n-k$, respectively. Correspondingly, suppose we partition U into $U = [U_1, U_2, U_3]$ with column sizes $r, s, m-r-s$, and V into $V = [V_1, V_2, V_3]$ with column sizes $p-k+r, s, k-r-s$, respectively. Observe that

$$X^T A^T A X = \begin{bmatrix} I_A & & & 0 \\ & S_A^2 & & \vdots \\ & & O_A^T O_A & \\ 0 & \ldots & & 0 \end{bmatrix},$$

$$X^T B^T B X = \begin{bmatrix} O_B^T O_B & & & 0 \\ & & S_B^2 & \vdots \\ & & & I_B \\ 0 & \ldots & & 0 \end{bmatrix},$$

where, for simplicity, we have used "0" to denote various zero matrices with appropriate sizes. Upon examining the second column block, we notice that

$$A^T A X_2 = B^T B X_2 \Lambda^2.$$

That is, $\{\lambda_i^2 | i = 1, \ldots, s\}$ is a subset of the eigenvalues of the symmetric pencil

$$A^T A - \mu B^T B. \tag{10}$$

Similarly, we point out the following remarks to include all other cases [3, 6].

1. If $k < n$, then $A^T A X_4 = B^T B X_4 = 0$ implies that *every* complex number is an eigenvalue of (10). This is the case that is considered of little interest. We will refer to eigenvalues of this type as *defective*.
2. Since $A^T A X_3 = 0$ and $B^T B X_3 \neq 0$, the pencil (10) has 0 as an eigenvalue with multiplicity $k-r-s$.
3. Since $B^T B X_1 = 0$ and $A^T A X_1 \neq 0$, we may regard that the pencil (10) has ∞ as an eigenvalue with multiplicity r.

We view the relationships

$$U_2^T A X_2 = S_A,$$
$$V_2^T B X_2 = S_B$$

as the fundamental and most important components of (9). We refer to the corresponding columns of U_2 and V_2 as the left generalized singular vectors of A and B, respectively. Note that there are two such left vectors for each generalized singular value, one for A, and one for B.

Similar to Theorem 1.1, we have the following variational formulation.

THEOREM 1.2. *Consider the optimization problem*

$$\max_{Bx \neq 0} \frac{\|Ax\|}{\|Bx\|}. \tag{11}$$

Then the generalized singular values $\lambda_1, \ldots, \lambda_s$ *of* (A, B) *are precisely the nonzero finite stationary values of the objective function in* (11).

Proof. The stationary values of $\|Ax\|/\|Bx\|$ are square roots of those of the function

$$f(x) := \frac{\langle Ax, Ax\rangle}{\langle Bx, Bx\rangle},$$

where $\langle\cdot,\cdot\rangle$ is the standard Euclidean inner product. It is not difficult to see that the gradient of f at x where $Bx \neq 0$ is given by

$$\nabla f(x) = \frac{2}{\langle Bx, Bx\rangle}\left(A^TAx - f(x)B^TBx\right).$$

The theorem follows from comparing the first-order condition $\nabla f(x) = 0$ with (10). ∎

Obviously, the corresponding stationary points $x \in R^n$ for the problem (11) are related to columns of the matrix X_2 (up to scalar multiplications), which are also eigenvectors of the pencil (10). What is not clear are the roles that U_2 and V_2 play in terms of the variational formula (11). In this note we present some new insights in this regard.

In the usual SVD duality theory the left singular vectors can be obtained from the optimization problem

$$\max_{y\neq 0} \frac{\|y^TA\|}{\|y\|}, \tag{12}$$

a formula similar to (2). Thus one might first guess that the duality theory analogous to (11) would be the problem

$$\max_{y^TB\neq 0} \frac{\|y^TA\|}{\|y^TB\|}.$$

However, this is certainly not a correct form as a single row vector y^T is not compatible for left multiplication on both A and B. We will see correct dual forms for the GSVD in (18) and (23).

2. Duality theory. For convenience, we denote

$$U_2 = [u_1^{(2)}, \ldots, u_s^{(2)}],$$
$$V_2 = [v_1^{(2)}, \ldots, v_s^{(2)}].$$

It follows from

$$U_2^TAX = S_AS_B^{-1}V_2^TBX = \Lambda V_2^TBX \tag{13}$$

that $U_2^TA = \Lambda V_2^TB$ or, equivalently,

$$C\begin{bmatrix} U_2 \\ -V_2\Lambda \end{bmatrix} = 0. \tag{14}$$

Note that the columns of both U_2 and V_2 are unit vectors. Given any $\begin{bmatrix} x \\ -y \end{bmatrix}$ in the null space of C with $\|x\| \neq 0$ and $\|y\| \neq 0$, we observe that

$$C\begin{bmatrix} \frac{x}{\|x\|} \\ -\frac{\|y\|}{\|x\|}\frac{y}{\|y\|} \end{bmatrix} = \frac{1}{\|x\|}C\begin{bmatrix} x \\ -y \end{bmatrix} = 0, \tag{15}$$

where $x/\|x\|$ and $y/\|y\|$ are also unit vectors. Comparing (15) with the relationship (14), we are motivated to consider the role that each generalized singular value λ_i plays in the optimization problem:

$$\max_{C\begin{bmatrix} x \\ -y \end{bmatrix}=0, x\neq 0} \frac{\|y\|}{\|x\|}. \tag{16}$$

However, we need to hastily point out a subtle flaw in the formulation of (16). Consider a given point $\begin{bmatrix} x \\ -y \end{bmatrix} \in \mathcal{S}$ with $\|x\| \neq 0$ and $\|y\| \neq 0$. Then $\begin{bmatrix} \alpha x \\ -\beta y \end{bmatrix} \in \mathcal{S}$ for arbitrary $\alpha, \beta \in R$. In this case, the optimization subproblem

$$\max_{x^T A = y^T B = 0, x\neq 0} \frac{\|y\|}{\|x\|} \tag{17}$$

becomes the problem

$$\max_{\alpha,\beta\in R, \alpha\neq 0} \frac{|\beta|}{|\alpha|}$$

that obviously has no stationary point at all and has maximum infinity. The trouble persists so long as $\begin{bmatrix} x \\ -y \end{bmatrix}$ contains components from $\mathcal{S}$. It is for this reason that the subspace $\mathcal{S}$ should be taken out of consideration. We should consider, instead of (16), the modified optimization problem (see (7) and (8))

$$\max_{\begin{bmatrix} x \\ -y \end{bmatrix} \in \mathcal{N}(Z), x\neq 0} \frac{\|y\|}{\|x\|}. \tag{18}$$

We will prove that each λ_i corresponds to a stationary value for the problem (18). But first it is worthy to point out some interesting remarks.

1. The optimization problem (18) is consistent with the ordinary singular value problem where $B = I$. In this case, $Z = \begin{bmatrix} A^T & I \\ N(A^T)^T & 0 \end{bmatrix}$. Thus $\begin{bmatrix} x \\ -y \end{bmatrix} \in \mathcal{N}(Z)$ implies that $y = A^T x \neq 0$. The forbidden situation $x^T A = y^T = 0$ in (18) is not a concern in this case because of the homogeneity in x and only implies that 0 is a stationary value (or equivalently A has a zero singular value). Thus in the case of the ordinary SVD, the problem (18) reduces to (12).
2. It is clear that $\dim(\mathcal{N}(C)) = m + p - k$ since we assume $\operatorname{rank}(C) = k$. The structure involved in (9) implies that for $\mathcal{S}$ defined in (4) it must be

$$\mathcal{S} = \mathcal{R}(U_3) \oplus \mathcal{R}(V_1).$$

That is, the size of $N(A^T)$ and $N(B^T)$ should be $m \times (m - r - s)$ and $p \times (p - k + r)$, respectively. It follows that $\dim(\mathcal{S}) = m + p - k - s$. The space we are interested in is the quotient space $\mathcal{N}(C)/\mathcal{S}$. It is known from the homomorphism theorem that $\dim(\mathcal{N}(C)/\mathcal{S}) = \dim(\mathcal{N}(C)) - \dim(\mathcal{S})$ [5, Lemma 4.8]. Thus $\dim(\mathcal{N}(C)/\mathcal{S}) = s$. We will see below that this dimension count agrees with our assumption that there are s generalized singular values.

The following theorem is critical to the study of stationary values and stationary points of the optimization problem (18).

THEOREM 2.1. *Let the columns of the matrix* $\begin{bmatrix} \Phi \\ \Psi \end{bmatrix}$ *with* $\Phi \in R^{m\times s}$ *and* $\Psi \in R^{p\times s}$ *be a basis for the subspace* $\mathcal{N}(Z)$. *Then the nondefective finite nonzero eigenvalues of the symmetric pencil of* (10),

$$A^TA - \mu B^TB,$$

are the same as those of the pencil

$$\Psi^T\Psi - \lambda\Phi^T\Phi. \tag{19}$$

Proof. Suppose $A^TAx = \mu B^TBx$. Since μ is nondefective and nonzero, $A^TAx = \mu B^TBx \neq 0$. That is, $\begin{bmatrix} Ax \\ -\mu Bx \end{bmatrix}$ represents a nonzero element in $\mathcal{N}(C)/\mathcal{S}$. Thus there exist vectors $v \in R^s$, $v \neq 0$, $\xi_A \in R^{m-r-s}$, $\xi_B \in R^{p-k+r}$ such that

$$\begin{aligned} Ax &= \Phi v + N(A^T)\xi_A, \\ -\mu Bx &= \Psi v + N(B^T)\xi_B. \end{aligned}$$

It follows that

$$(\Psi^T\Psi - \mu\Phi^T\Phi)v = -\mu(\Phi^TA + \Psi^TB)x + (\mu\Phi^TN(A^T)\xi_A - \Psi^TN(B^T)\xi_B) = 0.$$

In the above, we have used the fact that $Z\begin{bmatrix} \Phi \\ \Psi \end{bmatrix} = 0$. This shows that μ is an eigenvalue of (19) with v as the corresponding eigenvector.

To complete the eigenvalue (generalized singular value) set equality, suppose now that $(\Psi^T\Psi - \lambda\Phi^T\Phi)v = 0$ with $\lambda \neq 0, \infty$ and $v \neq 0$. We want to show that the equation

$$\begin{bmatrix} A \\ -\lambda B \end{bmatrix} x = \begin{bmatrix} \Phi v \\ \Psi v \end{bmatrix} \tag{20}$$

has a solution x. If this can be done, then since $[A^T, B^T]\begin{bmatrix} \Phi \\ \Psi \end{bmatrix} = 0$, it follows that x is an eigenvector of the pencil $A^TA - \mu B^TB$ with eigenvalue λ.

To show (20) means to show that the vector $\begin{bmatrix} \Phi v \\ \Psi v \end{bmatrix}$ is in the column space of the matrix $\begin{bmatrix} A \\ -\lambda B \end{bmatrix}$. It suffices to show that

$$\begin{bmatrix} \Phi v \\ \Psi v \end{bmatrix} \perp \begin{bmatrix} y \\ z \end{bmatrix} \tag{21}$$

wherever

$$\begin{bmatrix} A^T, & -\lambda B^T \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} = 0. \tag{22}$$

Rewrite (22) as $[A^T, B^T]\begin{bmatrix} y \\ -\lambda z \end{bmatrix} = 0$, showing that $\begin{bmatrix} y \\ -\lambda z \end{bmatrix} \in \mathcal{N}(C)$. We, therefore, must have

$$\begin{aligned} y &= \Phi w + N(A^T)\eta_A, \\ -\lambda z &= \Psi w + N(B^T)\eta_B \end{aligned}$$

for some vectors w, η_A, and η_B of appropriate size. Substituting y and z into (21) implies

$$\begin{bmatrix} y^T, & z^T \end{bmatrix} \begin{bmatrix} \Phi v \\ \Psi v \end{bmatrix} = w^T \left(\Phi^T \Phi v - \frac{1}{\lambda} \Psi^T \Psi v \right) + \left(\eta_A^T N(A^T)^T \Phi v - \frac{1}{\lambda} \eta_B N(B^T)^T \Psi v \right) = 0.$$

The assertion is therefore proved. ∎

COROLLARY 2.2. *The generalized singular values λ_i, $i = 1, \ldots, s$, are the stationary values associated with the optimization problem* (18).

Proof. We have already seen in Theorem 1.2 how the generalized singular values of (A, B) are related to the pencil $A^T A - \mu B^T B$, which are now related to the pencil $\Psi^T \Psi - \lambda \Phi^T \Phi$. By Theorem 1.2 again, we conclude that the generalized singular values of (A, B) can be found from the stationary values associated with the optimization problem

$$\max_{\Phi v \neq 0} \frac{\|\Psi v\|}{\|\Phi v\|}, \tag{23}$$

which is equivalent to (18). ∎

We now characterize the stationary points of (18). In particular, we prove the following result, which completes our duality theory. Aside from the fundamental connection between the GSVD and its duality theory, the eigenvalue *deflation* of the proof should be of special interest in its own right.

THEOREM 2.3. *Suppose*

$$\begin{bmatrix} x_1 \\ -y_1 \end{bmatrix} \cdots \begin{bmatrix} x_s \\ -y_s \end{bmatrix}$$

are stationary points for the problem (18) *with corresponding stationary values $\lambda_1, \ldots, \lambda_s$. Define*

$$u_i := \frac{x_i}{\|x_i\|}, \tag{24}$$

$$v_i := \frac{y_i}{\|y_i\|}. \tag{25}$$

Then the columns of the matrices $\tilde{U} := [u_1, \ldots, u_s]$ and $\tilde{V} := [v_1, \ldots, v_s]$ are the left generalized singular vectors of A and B, respectively.

Proof. Suppose $\begin{bmatrix} x_1 \\ -y_1 \end{bmatrix}$ is an associated stationary point of (18) with the stationary value λ_1. (The ordering of which stationary value is found is immaterial in the following discussion. We assume λ_1 is found first.) Taking this vector to be the first basis vector in $\mathcal{N}(Z)$, we may write

$$\begin{bmatrix} \Phi \\ \Psi \end{bmatrix} = \begin{bmatrix} x_1 & \Phi_2 \\ -y_1 & \Psi_2 \end{bmatrix},$$

where $\Phi_2 \in R^{m \times (s-1)}$ and $\Psi_2 \in R^{p \times (s-1)}$ are to be defined below. Consider the stacked matrix

$$Z_2 := \begin{bmatrix} A^T & B^T \\ N(A^T)^T & 0 \\ 0 & N(B^T)^T \\ x_1^T & 0 \end{bmatrix}.$$

Note that, due to the last row in Z_2, the null space of Z_2 is a proper subspace of the null space of Z with one less dimension. We may, therefore, use a basis of the null space of Z_2 to form the columns of the matrix $\begin{bmatrix} \Phi_2 \\ \Psi_2 \end{bmatrix}$. In this way, we attain the additional property that

$$x_1^T \Phi_2 = 0.$$

Note that the eigenvector of (19) corresponding to eigenvalue λ_1 is the same as the stationary point for the problem (23) with stationary value λ_1. Since (23) is simply a coordinate representation of (18) and we already assume that $\begin{bmatrix} x_1 \\ -y_1 \end{bmatrix}$ is a stationary point associated with (18), the eigenvector of (19) corresponding to λ_1 must be the unit vector $e_1 \in R^q$. It follows that

$$y_1^T \Psi_2 = 0,$$

and hence

$$\Psi^T \Psi - \lambda \Phi^T \Phi = \begin{bmatrix} y_1^T y_1 - \lambda x_1^T x_1 & 0 \\ 0 & \Psi_2^T \Psi_2 - \lambda \Phi_2^T \Phi_2 \end{bmatrix}.$$

Thus we have shown that the eigenvalues of the pencil $\Psi_2^T \Psi_2 - \lambda \Phi_2^T \Phi_2$ are exactly those of the pencil $\Psi^T \Psi - \lambda \Phi^T \Phi$ with λ_1 excluded. Note that the submatrix $\begin{bmatrix} \Phi_2 \\ \Psi_2 \end{bmatrix}$ spans a null subspace of Z that is complementary to the vector $\begin{bmatrix} x_1 \\ -y_1 \end{bmatrix}$. After the first stationary point is found, we may, therefore, deflate (18) to the problem

$$\max_{z_2 \begin{bmatrix} x \\ -y \end{bmatrix} = 0, x \neq 0} \frac{\|y\|}{\|x\|}. \tag{26}$$

A stationary point of (26) will also be a stationary point of (18) since it gives the same stationary value in both problems. This deflation procedure may be continued until all nonzero stationary values are found.

Then, by construction, $\tilde{U}^T \tilde{U} = I$ and $\tilde{V}^T \tilde{V} = I$. Furthermore, we have

$$\tilde{U}^T A = \Lambda \tilde{V}^T B,$$

which completes the proof. □

That is, we have derived two matrices $\tilde{U}$ and $\tilde{V}$ that play the same role as that of U_2 and V_2, in (9), respectively.

3. Summary. We have discussed a variational formulation for the GSVD of a pair of matrices. In particular, we characterize the role of the left generalized singular vectors in this formulation.

We summarize the analogies between the SVD and the GSVD in Table 1. The stationary values in any of the variational formulations give rise to the corresponding singular values.

There is a close correspondence between the (generalized) eigenvalue problem and the (generalized) singular value problem, as is indicated in Theorems 1.1 and 1.2. The results in Theorems 2.1 and 2.3 apparently are new and shed light on understanding the left singular vectors.

TABLE 1
Comparison of variational formulations between SVD and GSVD.

	Regular problem	Generalized problem
Decomposition	$U^T A = S V^T$	$U^T A X = \Lambda V^T B X$
	See formula (1)	See formula (13)
Right singular vector	V	X
Variational formula	$\max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$	$\max_{Bx \neq 0} \frac{\|Ax\|}{\|Bx\|}$
(including zero σ_i, λ_i)	See formula (2)	See formula (11)
Left singular vector	U	$[U^T, V^T]^T$
Variational formula	$\begin{bmatrix} A^T & \max I \\ N(A^T)^T & 0 \end{bmatrix} \begin{bmatrix} x \\ -y \end{bmatrix}_{=0, x \neq 0} \frac{\|y\|}{\|x\|}$	$\begin{bmatrix} A^T & \max B^T \\ N(A^T)^T & 0 \\ 0 & N(B^T)^T \end{bmatrix} \begin{bmatrix} x \\ -y \end{bmatrix}_{=0, x \neq 0} \frac{\|y\|}{\|x\|}$
	$\left(= {}_{A^T} \max_{x \neq 0, x \neq 0} \frac{\|A^T x\|}{\|x\|} \right)$	
(only positive σ_i, λ_i)	See formula (12)	See formula (18)

Some of the available numerical methods and approaches for computing the GSVD are available in [2, 7, 8, 10]. The deflation process used in the characterization of the left singular vectors can be carried out effectively by updating techniques [4]. We anticipate that the discussion here might lead to a new numerical algorithm, especially when a few singular values are required and the matrix C is sparse.

Acknowledgment. We want to thank an anonymous referee for the many valuable suggestions that significantly improved this paper.

REFERENCES

[1] B. ÅNSTRÖM, *The generalized singular value decomposition and* $(A - \lambda B)$*-problem*, BIT, 24 (1984), pp. 568–583.
[2] Z. BAI AND J. W. DEMMEL, *Computing the generalized singular value decomposition*, SIAM J. Sci. Comput., 14 (1993), pp. 1464–1486.
[3] B. DE MOOR AND G. H. GOLUB, *Generalized Singular Value Decomposition: A Proposal for a Standardized Nomenclature*, Manuscript NA-89-05, Stanford University, Stanford, CA, 1989.
[4] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., The John Hopkins University Press, Baltimore, MD, 1989.
[5] I. N. HERSTEIN, *Topics in Algebra*, 2nd ed., Xerox College Publishing, Lexington, MS, 1975.
[6] C. C. PAIGE AND M. A. SAUNDERS, *Towards a generalized singular value decomposition*, SIAM J. Numer. Anal., 18 (1981), pp. 398–405.
[7] C. C. PAIGE, *Computing the generalized singular value decomposition*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1126–1146.

[8] G. W. Stewart, *Computing the CS-decomposition of a partitioned orthonormal matrix*, Numer. Math., 40 (1982), pp. 297–306.
[9] C. F. Van Loan, *Generalizing the singular value decomposition*, SIAM J. Numer. Anal., 13 (1976), pp. 76–83.
[10] C. F. Van Loan, *Computing the CS and the generalized singular value decomposition*, Numer. Math., 46 (1985), pp. 479–491.

SIAM J. MATRIX ANAL APPL.
Vol. 18, No. 4, pp. 1093–1095, October 1997

018

INEQUALITIES FOR THE SINGULAR VALUES OF HADAMARD PRODUCTS*

XINGZHI ZHAN†

Abstract. Let $M_{m,n}$ be the space of $m \times n$ complex matrices. For $A, B \in M_{m,n}$, denote the Hadamard (or Schur) product of A and B by $A \circ B$. Given $A \in M_{m,n}$, let $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_{\min\{m,n\}}(A)$ be the ordered singular values, and the decreasingly ordered Euclidean row and column lengths of A are denoted by $r_1(A) \geq r_2(A) \geq \cdots \geq r_m(A)$ and $c_1(A) \geq c_2(A) \geq \cdots \geq c_n(A)$, respectively. It is shown that for any $A, B \in M_{m,n}$,

$$\sum_{i=1}^{k} \sigma_i(A \circ B) \leq \sum_{i=1}^{k} \min\{c_i(A), r_i(A)\}\sigma_i(B),$$
$$k = 1, 2, \ldots, \min\{m, n\}.$$

This settles, in a stronger form, a conjecture of R. A. Horn and C. R. Johnson [*Topics in Matrix Analysis*, Cambridge University Press, New York, 1991, p. 344] affirmatively.

Key words. singular values, Hadamard products

AMS subject classifications. 15A18, 15A42, 15A45

PII. S0895479896309645

1. Introduction.

Let $M_{m,n}$ be the space of $m \times n$ complex matrices and $M_n \equiv M_{n,n}$. For $A = [a_{ij}], B = [b_{ij}] \in M_{m,n}$, the Hadamard product of A and B is $A \circ B \equiv [a_{ij}b_{ij}]$. Much work has been done for the singular values of Hadamard products. See [2] and [3].

We always arrange the singular values of $A \in M_{m,n}$ in decreasing order $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_{\min\{m,n\}}(A)$. Denote the decreasingly ordered Euclidean row and column lengths of $A \in M_{m,n}$ by $r_1(A) \geq r_2(A) \geq \cdots \geq r_m(A)$ and $c_1(A) \geq c_2(A) \geq \cdots \geq c_n(A)$, respectively; i.e., $r_k(A)$ is the kth largest value of $(\sum_{j=1}^{n} |a_{ij}|^2)^{1/2}, i = 1, \ldots, m$ and $c_k(A)$ is the kth largest value of $(\sum_{i=1}^{m} |a_{ij}|^2)^{1/2}, j = 1, \ldots, n$. In [3, p. 344] R. A. Horn and C. R. Johnson ask whether inequalities of the form

$$\sum_{i=1}^{k} \sigma_i(A \circ B) \leq \sum_{i=1}^{k} c_i(A)^{\alpha} r_i(A)^{1-\alpha} \sigma_i(B) \tag{1}$$

are valid for $0 \leq \alpha \leq 1$ and $k = 1, \ldots, \min\{m, n\}$. The only values of α for which (1) has been proved are $\alpha = 0, 1/2$, and 1 [3, Theorems 5.5.20 and 5.5.21].

In this paper we shall prove a stronger result than (1).

2. The result.

THEOREM 1. *For any $A, B \in M_{m,n}$,*

$$\sum_{i=1}^{k} \sigma_i(A \circ B) \leq \sum_{i=1}^{k} \min\{c_i(A), r_i(A)\}\sigma_i(B) \tag{2}$$

*Received by the editors September 20, 1996; accepted for publication (in revised form) by T. Ando November 22, 1996.

http://www.siam.org/journals/simax/18-4/30964.html

†Institute of Mathematics, Peking University, Beijing 100871, China (zhan@sxx0.math.pku.edu.cn).

$$k = 1, 2, \ldots, \min\{m, n\}.$$

We need consider only the square case $m = n$, since nonsquare matrices can be augmented to square ones with zero blocks. For Hermitian matrices $H, G \in M_n$ we write $H \le G$ to mean that $G - H$ is positive semidefinite. Denote by I the identity matrix. The key observation is the following fact.

LEMMA 2. *For any $A, B \in M_n$ we have*
(a) $(A \circ B)(A \circ B)^* \le \sigma_1(B)^2 I \circ (AA^*)$,
(b) $(A \circ B)^*(A \circ B) \le \sigma_1(B)^2 I \circ (A^*A)$, *and*
(c) $\sigma_i(A \circ B) \le \min\{c_i(A), r_i(A)\}\sigma_1(B)$ *for* $i = 1, 2, \ldots, n$.

Proof. It is known [2, p. 116] that

$$(A \circ B)(A \circ B)^* \le (AA^*) \circ (BB^*) \tag{3}$$

for all $A, B \in M_n$. Since $BB^* \le \sigma_1(B)^2 I$, the Schur product theorem implies

$$(AA^*) \circ (BB^*) \le (AA^*) \circ (\sigma_1(B)^2 I). \tag{4}$$

Combining (3) with (4) yields (a). Now replace A and B in (a) by their adjoints to get (b). Note that $0 \le X \le Y \Longrightarrow \sigma_i(X) \le \sigma_i(Y)$ $(i = 1, 2, \ldots)$ [5, Corollary 7.7.4(c)]. By the definition of $r_i(A)$ and $c_i(A)$, (c) follows from (a) and (b). □

The next result is a summary of some ideas in [1] and [4], from which we may establish the main theorem.

LEMMA 3. *Let $A \in M_n$ and $\alpha_1 \ge \alpha_2 \ge \cdots \ge \alpha_n \ge 0$ be given, and suppose*

$$\sum_{i=1}^{k} \sigma_i(A \circ B) \le \sum_{i=1}^{k} \alpha_i \sigma_1(B),\ k = 1, \ldots, n$$

for all $B \in M_n$. Then

$$\sum_{i=1}^{k} \sigma_i(A \circ B) \le \sum_{i=1}^{k} \alpha_i \sigma_i(B),\ k = 1, \ldots, n$$

for all $B \in M_n$.

Proof. A matrix $K \in M_n$ is called a *rank r partial isometry* if $\sigma_1(K) = \cdots = \sigma_r(K) = 1$ and $\sigma_{r+1}(K) = \cdots = \sigma_n(K) = 0$. The proof of Lemma 8 in [1] with $c_i(X)c_i(Y)$ replaced there by α_i shows that $|\mathrm{tr}[(A \circ K_r)K_s]| \le \sum_{i=1}^{\min\{r,s\}} \alpha_i$ for any partial isometries $K_r, K_s \in M_n$ with respective ranks r and s. Next the argument following Lemma 10 in [1] completes the proof of this lemma. □

Proof of Theorem 1. We need only prove the square case $m = n$. Lemma 2(c) implies $\sum_{i=1}^{k} \sigma_i(A \circ B) \le \sum_{i=1}^{k} \min\{c_i(A), r_i(A)\}\sigma_1(B)$ for $k = 1, \ldots, n$. Applying Lemma 3 completes the proof. □

Since for any $0 \le \alpha \le 1$, $\min\{c_i(A), r_i(A)\} \le c_i(A)^\alpha r_i(A)^{1-\alpha}$, (2) is evidently sharper than (1).

Finally we remark that the weak multiplicative majorization analogue of (2), which is stronger than (2), is, in general, false. Consider the following example:

$$A = I_2 \quad \text{and} \quad B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Acknowledgments. The author wishes to thank Professor T. Ando for his encouragement and the three referees for their helpful suggestions which make the paper more succinct.

REFERENCES

[1] T. Ando, R. A. Horn, and C. R. Johnson, *The singular values of a Hadamard product: A basic inequality*, Linear and Multilinear Algebra, 21 (1987), pp. 345–365.

[2] R. A. Horn, *The Hadamard product*, in Matrix Theory and Applications, Proceedings of Applied Mathematics, Vol. 40, C. R. Johnson, ed., AMS, Providence, RI, 1990.

[3] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[4] R. A. Horn and C. R. Johnson, *Hadamard and conventional submultiplicativity for unitarily invariant norms on matrices*, Linear and Multilinear Algebra, 20 (1987), pp. 91–106.

[5] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.